



**HAL**  
open science

# Système à base de connaissances pour le processus de plan d'expériences numériques

Gaëtan Blondet

► **To cite this version:**

Gaëtan Blondet. Système à base de connaissances pour le processus de plan d'expériences numériques. Mécanique [physics.med-ph]. Université de Technologie de Compiègne, 2017. Français. NNT : 2017COMP2363 . tel-01693875

**HAL Id: tel-01693875**

**<https://theses.hal.science/tel-01693875v1>**

Submitted on 26 Jan 2018

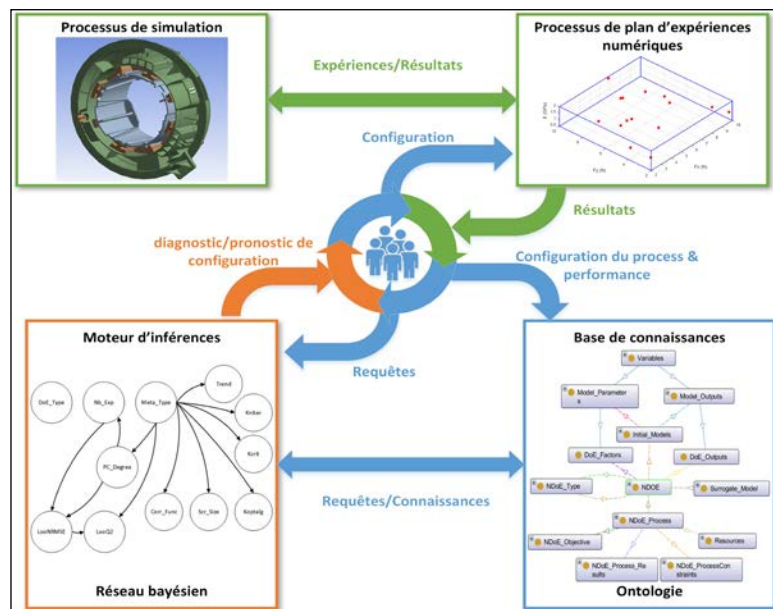
**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Par **Gaëtan BLONDET**

*Système à base de connaissances pour le processus de plan d'expériences numériques*

Thèse présentée  
 pour l'obtention du grade  
 de Docteur de l'UTC



Soutenue le 9 juin 2017

**Spécialité** : Ingénierie Industrielle : Unité de recherche en Mécanique - Laboratoire Roberval (UMR-7337)

D2363



## THÈSE

pour obtenir le grade de docteur délivré par

**Sorbonne Universités**

**Université de Technologie de Compiègne**

**Spécialité “Ingénierie industrielle”**

École doctorale n° 71 : Sciences pour l'ingénieur  
Laboratoire Roberval (CNRS, UMR 7337)

Unité de Recherche en Mécanique  
*présentée et soutenue publiquement par*

**Gaëtan BLONDET**

le 9 Juin 2017

# **Systeme à base de connaissances pour le processus de plan d'expériences numériques**

### Jury

M. Abdelaziz Bouras, Professeur des Universités, Qatar University,	Rapporteur
M. Bertrand Rose, Maître de Conférences HDR, Université de Strasbourg,	Rapporteur
Mme Michelle Sergent, Professeur des Universités, Université Aix-Marseille,	Examineur
Mme Zorha Cherfi-Boulanger, Professeur des Universités, Université de Technologie de Compiègne,	Examineur
M. Alain Bernard, Professeur des Universités, Ecole Centrale de Nantes,	Examineur
M. Gilles Besombes, Responsable Calcul Mécanique, Valeo,	Membre invité
M. Julien Le Duigou, Enseignant-Chercheur, Université de Technologie de Compiègne,	Directeur de Thèse
M. Nassim Boudaoud, Maître de Conférences, Université de Technologie de Compiègne,	Directeur de Thèse



# Table des matières

Liste des figures	iii
Liste des tableaux	v
Remerciements	vii
Introduction	1
<b>1 Contexte industriel et problématique</b>	<b>5</b>
1.1 Contexte des travaux de recherche	8
1.1.1 Contexte industriel	8
1.1.2 Gestion du cycle de vie du produit	12
1.1.3 Simulation numérique	15
1.1.4 Prise en compte des variabilités	18
1.2 Processus de plan d'expériences numériques	21
1.2.1 Définition d'un plan d'expériences	22
1.2.2 Objectifs des plans d'expériences numériques	24
1.2.3 Définition d'un processus de plan d'expériences numériques	24
1.2.4 Types de plan d'expériences	26
1.2.5 Méthodes d'estimation des modèles de substitution	30
1.2.6 Méthodes d'analyse de sensibilité	37
1.2.7 Algorithmes d'optimisation	40
1.2.8 Synthèse : Le processus de plan d'expériences numériques	41
1.3 Gestion des données pour la simulation numérique	43
1.3.1 Gestion du cycle de vie des simulations	43
1.3.2 Gestion des données de simulation	45
1.3.3 Synthèse sur la gestion des données de simulation	45
1.4 Conclusion et problématique	47
<b>2 État de l'art</b>	<b>49</b>
2.1 Gestion des connaissances	53
2.2 Modèles de données et de connaissances	58
2.2.1 Modèles de données	58
2.2.2 Modèles de connaissances	65
2.2.3 Synthèse sur les modèles de connaissances	72
2.3 Réutilisation des connaissances	74
2.3.1 Critères d'analyses	76
2.3.2 Méthodes de réutilisation	77
2.3.3 Comparaison des méthodes de réutilisation	85
2.4 Synthèse	86

<b>3</b>	<b>Proposition : système à base de connaissances pour l'aide à la configuration des processus de plan d'expériences numériques</b>	<b>87</b>
3.1	Architecture globale du système . . . . .	91
3.2	Ontologie du domaine des processus de plan d'expériences numériques . . . . .	94
3.2.1	Objectifs principaux . . . . .	94
3.2.2	Description de l'ontologie pour les processus de plan d'expériences numériques . . . . .	94
3.2.3	Retour sur les critères de conception de l'ontologie . . . . .	100
3.2.4	Utilisation de l'ontologie . . . . .	106
3.2.5	Apports de l'ontologie . . . . .	109
3.3	Réutilisation des connaissances capitalisées par réseaux bayésiens . . . . .	110
3.3.1	Conception du réseau bayésien . . . . .	110
3.3.2	Validation des performances prédictives du réseau bayésien . . . . .	119
3.3.3	Apports du réseau bayésien . . . . .	124
3.4	Synthèse et discussion . . . . .	126
<b>4</b>	<b>Application du système proposé</b>	<b>129</b>
4.1	Implémentation . . . . .	133
4.1.1	Outils utilisés . . . . .	133
4.1.2	Modèles de simulation numérique utilisés . . . . .	135
4.1.3	Instanciation de la base de connaissances . . . . .	138
4.2	Exemple d'utilisation . . . . .	141
4.2.1	Situation initiale . . . . .	141
4.2.2	Extraction des connaissances capitalisées . . . . .	144
4.2.3	Apprentissage du réseau bayésien . . . . .	146
4.2.4	Estimation des performances prédictives du réseau bayésien . . . . .	153
4.2.5	Inférences du réseau bayésien : prédiction de la performance de la configuration de processus de plan d'expériences numériques . . . . .	156
4.2.6	Inférences du réseau bayésien : prédiction de configurations performantes de processus de plan d'expériences numériques . . . . .	156
4.2.7	Vérification et exécution de la configuration proposée . . . . .	160
4.3	Conclusion de l'application . . . . .	162
	<b>Conclusion</b>	<b>165</b>
	<b>Bibliographie</b>	<b>171</b>
	<b>A Figures annexes</b>	<b>I</b>

# Liste des figures

1.1	Plan du chapitre 1	7
1.2	Cycle de vie du produit réel	13
1.3	Cycle de vie du produit virtuel	13
1.4	Le processus de simulation numérique	16
1.5	Représentation 3D de la structure du modèle T1	21
1.6	Définition d'un plan d'expériences numériques pour le modèle T1.	23
1.7	Le processus de PEN	25
1.8	Exemple d'un PEN de type Halton	28
1.9	Processus de PEN adaptatif	29
1.10	Processus d'apprentissage pour la production d'un modèle de substitution	33
1.11	Surface de réponses du modèle T1	34
1.12	Exemple de Perceptron Multi-Couches	35
1.13	Synthèse des méthodes d'analyse de sensibilité	38
1.14	Indices de Sobol' pour le modèle T1	39
2.1	Plan du chapitre 2	52
2.2	Cycle synthétique de la gestion de la connaissances	54
2.3	Modèle de conversion des connaissances	55
2.4	Structure globale d'un système d'aide à la décision à base de connaissances	57
2.5	Fonctionnalités de l'AP242 édition 2	60
2.6	Architecture de l'AP 209 édition 2	61
2.7	Classification synthétique des modèles de données étudiées	64
2.8	Standards du web spécifiés par le W3C	66
2.9	Exemple de relations ontologiques	66
2.10	Classification synthétique des ontologies étudiées	73
2.11	Principaux domaines de l'intelligence artificielle	75
2.12	Exemple d'arbre de décision	81
2.13	Exemple de graphe d'un réseau bayésien	82
2.14	Fonctionnement du raisonnement à base de cas	83
3.1	Plan du chapitre 3	90
3.2	Architecture globale du système à base de connaissances proposé	91
3.3	Caractéristiques de ODE	95
3.4	Arborescence de ODE	95
3.5	Extrait de ODE (PEN)	97
3.6	Extrait de ODE (PEN adaptatif)	98
3.7	Détails des données de la classe Kriging.	99
3.8	Vérification de l'ontologie	101
3.9	Extrait de ODE (Optimisation)	101
3.10	Représentation des liens inclus dans ODE	104

3.11	Détails de la classe "NDoE_Type" et lien avec l'ontologie STATO . . . . .	104
3.12	Extrait de ODE : annotations . . . . .	105
3.13	Instance de la classe NDOE process . . . . .	106
3.14	Instance de la classe NDOE . . . . .	106
3.15	Résultat de la requête SPARQL . . . . .	108
3.16	Evolution de la structure du réseau bayésien pour le modèle T1 en PEN adaptatif . . . . .	116
3.17	Différence d'estimation des paramètres du réseau bayésien entre le maximum de vraisemblance et la méthode bayésienne . . . . .	118
3.18	Validation croisée : Erreur globale "k-fold" . . . . .	120
3.19	Histogramme des indicateurs de prédictivité du réseau bayésien . . . . .	122
3.20	Architecture du système à base de connaissances proposé . . . . .	126
4.1	Plan du chapitre 4 . . . . .	132
4.2	Outils utilisés pour le système à base de connaissances . . . . .	134
4.3	Logigramme du système à base de connaissances . . . . .	136
4.4	Représentation 3D de la structure des modèles T . . . . .	137
4.5	Système de climatisation . . . . .	137
4.6	Représentation 3D du système de découplage du pulseur . . . . .	137
4.7	Modélisation du système de découplage pour simulation . . . . .	137
4.8	IHM : fenêtre d'initialisation du processus de PEN . . . . .	141
4.9	IHM : fenêtre de configuration initial du processus de PEN . . . . .	142
4.10	Composition des ensembles d'apprentissage et de validation (1/2) . . . . .	148
4.11	Composition des ensembles d'apprentissage et de validation (2/2) . . . . .	149
4.12	Graphe du réseau bayésien après apprentissage . . . . .	150
4.13	Probabilités conditionnelles estimées par inférence bayésienne . . . . .	152
4.14	IHM : fenêtre de rapport des performances prédictives du réseau bayésien . . . . .	154
4.15	IHM : fenêtre de prédiction des performances d'une configuration imposée . . . . .	157
4.16	Liste des configurations proposées par le système d'aide à la décision . . . . .	158
4.17	IHM : fenêtre de prédiction des configurations performantes . . . . .	159
4.18	Exécution de la nouvelle configuration issue de l'utilisation du système d'aide à la décision. . . . .	161
4.19	IHM : fenêtre des résultats du système d'aide à la configuration du processus de PEN . . . . .	162
A.1	Logigramme du processus de PEN implémenté . . . . .	II
A.2	Logigramme du processus de PEN implémenté pour la production de modèle de substitution . . . . .	III



# Liste des tableaux

1.1	Caractéristiques principales du modèle T1 . . . . .	22
2.1	Exemples de caractéristiques d'une propriété objet : analyse pour le cas des concepts de PEN et de facteurs . . . . .	67
2.2	Tableau comparatif des méthodes d'analyse des connaissances capitalisées . . . . .	85
3.1	Matrice de confusion réseau bayésien . . . . .	120
3.2	Matrice de confusion et indicateurs de performance prédictive du réseau bayésien . . . . .	123
4.1	Caractéristiques principales des modèles numériques utilisés . . . . .	138
4.2	Définition des ensembles de données générés pour la base de connaissances. . . . .	139
4.3	Variables et modalités des processus de PEN considérés . . . . .	139
4.4	Variables et modalités des processus de PEN adaptatif considérés . . . . .	140
4.5	Caractéristiques principales de la configuration complète du processus de PEN proposée par l'utilisateur . . . . .	143
4.6	Résultat de la requête SPARQL n°1 . . . . .	145
4.7	Résultat de la requête SPARQL n°2 . . . . .	145
4.8	Matrice de confusion et indicateurs de performance prédictive du réseau bayésien . . . . .	155
4.9	Caractéristiques de la nouvelle configuration du processus de PEN . . . . .	160



# Remerciements

Les quelques deux cent pages que vous vous apprêtez à lire, sans doute avec joie et entrain, n'auraient jamais pu être écrites sans le soutien de nombreuses personnes, avec lesquelles j'ai eu un grand plaisir à travailler et à vivre.

Je tiens à remercier en premier lieu mes directeurs de thèse, Julien Le Duigou et Nassim Boudaoud, pour leur aide, leurs conseils avisés et leur appui tout au long de ces trois années. Je remercie également Benoît Eynard pour le soutien qu'il m'a accordé.

Je remercie l'ensemble du jury pour l'intérêt qu'ils ont porté à mes travaux et aux échanges très enrichissants que nous avons pu avoir. Un grand merci à Michelle Sergent, pour avoir présidé ce jury ; à Bertrand Rose et à Abdelaziz Bouras pour avoir accepté d'être les rapporteurs de ma thèse ; à Alain Bernard et Zorha Cherfi-Boulanger pour avoir examiné mes travaux et pour leurs remarques très pertinentes.

Également présent au sein de ce jury, je tiens à adresser un remerciement tout particulier à Gilles Besombes, pour avoir suivie mes travaux tout au long de ma thèse avec grand intérêt. De même, je tiens à remercier tous les autres partenaires du projet SDM4DOE, financé par le Fond Unique Interministériel et labellisé par le pôle de compétitivité Systematic. Je remercie en particulier Farouk Belkadi, pour tous les riches échanges que nous avons eus.

J'adresse de vifs remerciements à Paul Schimmerling et à tous les membres du groupe de travail "plans d'expériences" pour l'AFNOR. Ils m'ont apporté, en plus de leur expertise scientifique et technique, de nombreuses opportunités très enrichissantes.

Un grand merci à toute l'équipe du bâtiment I. Je n'oublierai jamais tous ces repas, et notamment les grandes discussions du vendredi, qui survenaient parfois le mardi. Je n'oublierai pas non plus tous les conseils, rires et coups de mains qu'ils m'ont apportés. Un grand merci, donc, à Emmanuel -notre mentor à tous-, à Philippe -son principal rival-, à Mathieu, Jean-Luc, Etienne, Alex, Amélie, Charles, Sylvie, Muriel, Magali, Christine, Frédéric, Laurent,... Un merci tout spécial pour Joanna, pour sa bonne humeur et ses excellents conseils !

Encore merci à tous les doctorants et post-doctorants que j'ai eu la chance de rencontrer : Christophe, Marianne - ma sœur d'armes-, Marina, Fabien, Mohammed, Cong, John, Julia, Emeric, Arnaud -comté power!-, Zoubida, Omar, Mounir, Jinhua, Pierre et Sophie, Doriane, Sophie, Christina, et aussi à tous les membres du RED<sup>2</sup> ! Merci à toutes celles et ceux que j'ai pu rencontrer durant mes études et ailleurs, et qui m'ont fait avancer.

J'ai bien sûr une pensée pour tous les autres, qui ont tellement compté : Fanny -best fillote ever!-, Coco, Pum, Nanard, Juju, Eric, Jonathan, Emilien, Liza. Pour finir, je ne remerciais jamais assez ma famille, parents, frères, grands-parents, neveux et nièces, et surtout mon filleul... pour tout leur soutien !

Il ne vous reste plus que 186 pages à lire, alors préparer la tisane, éteignez votre télé et bon voyage au pays des connaissances et des probabilités !

# Introduction



La compétitivité des entreprises de l'industrie mécanique est confrontée à un contexte concurrentiel mondialisé et fortement collaboratif. Le développement de produits doit être plus rapide, moins coûteux, et aboutir à des produits de meilleure qualité, prenant en compte les exigences des clients et des réglementations.

La grande progression des outils et moyens numériques de ces dernières décennies a offert des moyens performants pour réduire les temps de développement, réduire les coûts, optimiser les produits et gérer des processus de plus en plus collaboratifs au sein d'entreprises étendues. D'une part, les outils de simulation numérique permettent aux bureaux d'études de simuler très précisément le comportement des produits et des processus de fabrication complexes. L'émergence des simulations multi-physiques, des optimisations multi-objectifs et des plans d'expériences numériques a grandement contribué à maintenir la compétitivité et l'innovation au sein des entreprises. D'autre part, les technologies de communication ont progressé au point de pouvoir échanger des informations rapidement, en très larges volumes, sur de grandes distances et depuis de nombreux sites différents. Le maintien des performances de l'entreprise et le respect des réglementations rendent de plus en plus nécessaire une gestion des informations utilisées par l'entreprise tout au long du cycle de vie du produit. La capitalisation des informations assurent une traçabilité et une capacité d'amélioration cruciale pour l'entreprise.

L'utilisation de techniques de plans d'expériences numériques apporte des moyens de rendre le produit plus robuste, plus fiable et mieux optimisé, très tôt dans le développement du produit. Cela mène à une réduction des essais physiques sur des prototypes coûteux. Ces techniques de plan d'expériences numériques s'avèrent tout de même coûteuses pour des produits complexes. Elles sont également difficiles à appliquer et mènent à une forte croissance des données à gérer dans l'entreprise.

Le projet FUI16 SDM4DOE (Simulation Data Management for Design of Experiment) a débuté en 2014 dans l'objectif de lever ces verrous liés à l'application des plans d'expériences numériques et pouvoir proposer leur application dans tous les secteurs de l'industrie mécanique. Les travaux de recherches présentés dans ce manuscrit ont été réalisés en collaboration avec le consortium d'entreprises et de laboratoires de recherches de ce projet, présentant en premier lieu des intérêts dans les secteurs de l'automobile et du génie civil. Ce manuscrit présente l'état d'avancement des recherches menées dans ce contexte pour proposer une solution pour favoriser l'utilisation des plans d'expériences numériques dans les bureaux d'études par la gestion des connaissances de l'entreprise.

Ce manuscrit est composé de quatre chapitres :

- le chapitre 1 présente les éléments de contexte industriel et les verrous scientifiques actuels concernant les activités de simulation numérique, de plans d'expériences numériques et de gestion des données de simulations. A la suite de ces analyses, la

problématique de ces travaux est énoncée pour cibler les solutions pouvant améliorer l'applicabilité des méthodes de plans d'expériences à des cas industriels ;

- le chapitre 2 présente l'état de l'art réalisé sur la base de la problématique de recherche. Les solutions de structuration et de réutilisation des informations sont analysées pour identifier les manques et définir les apports nécessaires pour répondre à la problématique ;
- le chapitre 3 détaille la proposition énoncée en réponse aux manques relevés. D'une part, une structuration des connaissances par un modèle ontologique spécifique est proposée, pour assurer la capitalisation, le partage et la réutilisation des données, connaissances et décisions liés aux processus de plans d'expériences numériques. D'autre part, se basant sur la première proposition, un système d'aide à la décision capable de réutiliser les données et les connaissances capitalisées est proposé. Ce système a pour objectif d'assister les concepteurs dans leurs choix méthodologiques et de proposer de nouvelles méthodes innovantes ;
- le chapitre 4 met en application la proposition globale sur un cas industriel issu de l'industrie automobile, afin d'illustrer le fonctionnement du système à base de connaissances et de montrer ses capacités à lever les verrous industriels et scientifiques identifiés.

La conclusion de ce manuscrit établira un bilan des recherches réalisées au cours de ces trois années. Les apports de ces travaux et les évolutions possibles de ce système sont présentés.



# Chapitre 1

## Contexte industriel et problématique

■ *L'expérience est la somme des premières fois...* ■

---

R.Hennegrave

### Sommaire

---

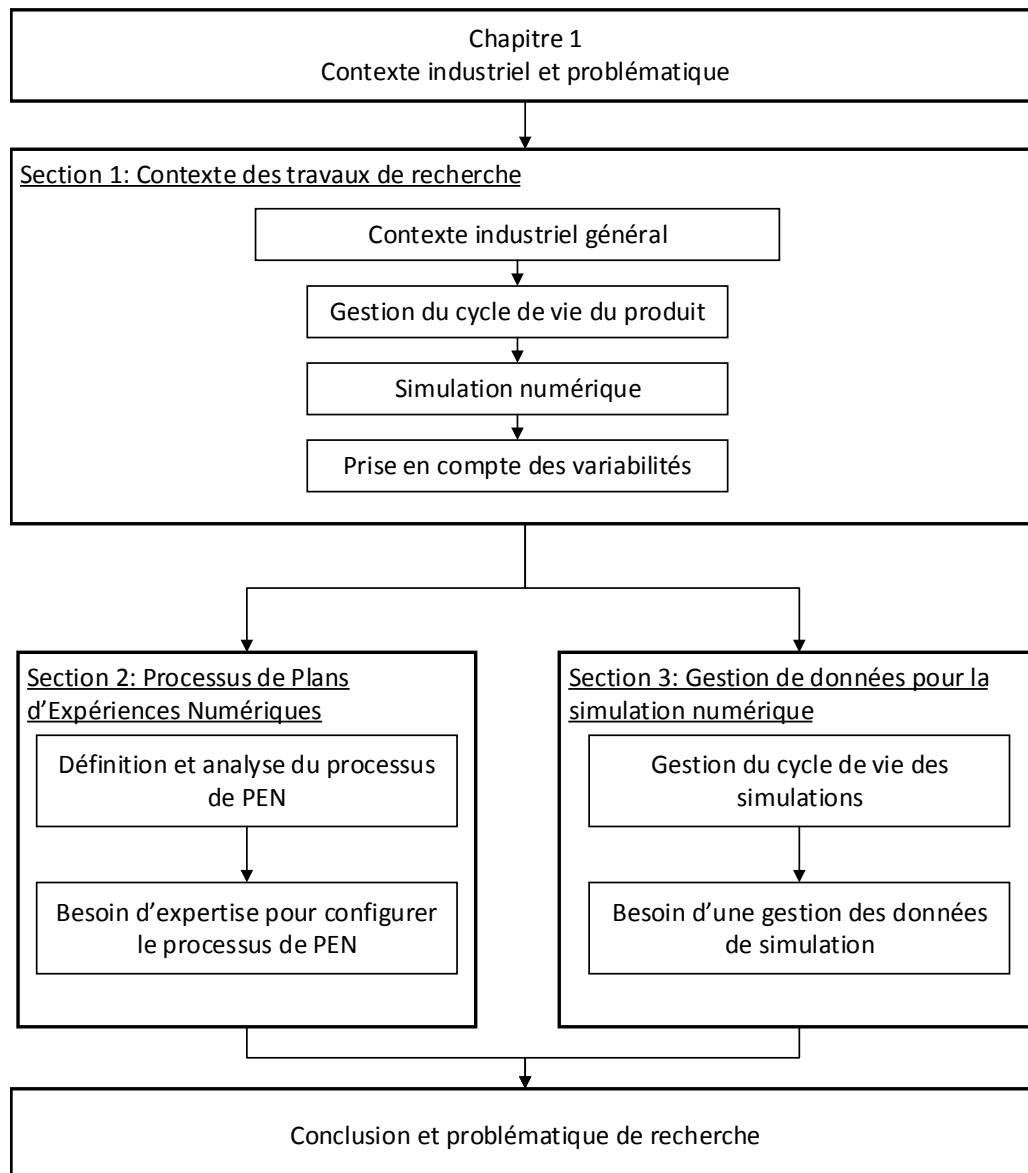
<b>1.1</b>	<b>Contexte des travaux de recherche</b>	<b>8</b>
1.1.1	Contexte industriel	8
1.1.2	Gestion du cycle de vie du produit	12
1.1.3	Simulation numérique	15
1.1.4	Prise en compte des variabilités	18
<b>1.2</b>	<b>Processus de plan d'expériences numériques</b>	<b>21</b>
1.2.1	Définition d'un plan d'expériences	22
1.2.2	Objectifs des plans d'expériences numériques	24
1.2.3	Définition d'un processus de plan d'expériences numériques	24
1.2.4	Types de plan d'expériences	26
1.2.5	Méthodes d'estimation des modèles de substitution	30
1.2.6	Méthodes d'analyse de sensibilité	37
1.2.7	Algorithmes d'optimisation	40
1.2.8	Synthèse : Le processus de plan d'expériences numériques	41
<b>1.3</b>	<b>Gestion des données pour la simulation numérique</b>	<b>43</b>
1.3.1	Gestion du cycle de vie des simulations	43
1.3.2	Gestion des données de simulation	45
1.3.3	Synthèse sur la gestion des données de simulation	45
<b>1.4</b>	<b>Conclusion et problématique</b>	<b>47</b>

---



Ce premier chapitre définit le contexte des recherches réalisées dans le cadre de cette thèse (Figure 1.1). La section 1 présente le contexte industriel et scientifique, centré sur le domaine de la mécanique et plus particulièrement sur les problématiques numériques. La section 2 traite spécifiquement d'une solution envisagée pour répondre aux problématiques industrielles : les plans d'expériences numériques ([Plans d'Expériences Numériques \(PEN\)](#)). Les [PEN](#) étendent l'usage de la simulation numérique pour la prise en compte des variabilités liées aux produits. Le processus de [PEN](#) est analysé et ses limites sont discutées, notamment concernant la génération massive de données de simulation et la complexité de mise en place d'un tel processus. La section 3 se focalise sur les aspects de gestion de données de simulation numérique. Ces deux dernières sections convergent vers la problématique de ces travaux, présentée en section 4.

FIGURE 1.1 – Plan du chapitre 1



## 1.1 Contexte des travaux de recherche

### 1.1.1 Contexte industriel

Les recherches menées dans le cadre de cette thèse ont pour périmètre l'industrie mécanique. L'industrie mécanique occupe une place importante dans de nombreux secteurs, tels que les secteurs de l'automobile, du ferroviaire, de l'aéronautique, du spatial ou encore de l'énergie. Afin de maintenir leur compétitivité, les entreprises de l'industrie mécanique font face à une nécessité d'innover, d'améliorer la qualité de leurs produits, de diminuer le temps de mise sur le marché et de diminuer également le coût de leurs produits.

Le livre blanc de la recherche en mécanique [AFM 15] présente, entre autres, les enjeux des différents secteurs. Ces enjeux concernent notamment la réduction de l'impact environnemental, l'amélioration de la sécurité, l'utilisation d'outils de simulations numériques pour des produits complexes, la prise en compte des variabilités et des incertitudes, la gestion des données et des connaissances, le développement collaboratif de produits, ou encore la mise en place de réseaux de capteurs pour le suivi des produits.

Dans le secteur automobile, la simulation numérique permet d'aboutir à une réduction de la masse du véhicule, à l'amélioration de son aérodynamisme ou encore à l'optimisation du rendement énergétique de la motorisation. La réduction de la masse passe par la simulation du comportement du produit, mais aussi par la simulation des procédés de fabrication et d'assemblage innovants. La simulation des écoulements de l'air autour des véhicules permet de réduire les forces de traînée, et donc les pertes d'énergies. Ce type de simulation tend à être utilisé de façon plus précise. L'amélioration de la motorisation passe, à court terme, par la modélisation multi-physique (fluide-structure) et l'optimisation. Ces trois axes participent à l'amélioration de l'efficacité énergétique du véhicule. De plus, la prise en compte des incertitudes liées aux produits et aux moyens de productions constitue un axe d'amélioration des activités de simulation. D'autres aspects de la mécanique sont concernés, tels que l'éco-conception, l'éco-fabrication, les sciences des matériaux ou encore la réalité virtuelle.

L'amélioration de l'efficacité énergétique concerne également les secteurs de l'aéronautique et du spatial. Le programme Vision 2050, défini par L'IATA (International Air Transport Association) en 2011 [IATA 11], présente des objectifs de réduction de l'impact environnemental en agissant sur le trafic aérien et sur les performances des avions. La volonté de réduire les émissions de gaz nocifs (Dioxyde de carbone : -75% ; oxydes d'azote : -90%) et la pollution sonore (bruit perçu : -65%) nécessite une optimisation de la masse, de la motorisation et de l'aérodynamisme des appareils. Ce secteur tend à placer les activités de simulation en amont des phases de développements pour minimiser le nombre d'essais et obtenir des modèles prédictifs au plus tôt. De plus, les essais nécessitent des améliorations

en termes de gestion des données pour soutenir les étapes de certifications des appareils. La capitalisation des connaissances et l'ingénierie concourantes sont aussi identifiées comme étant des axes de développement.

Les industries de l'énergie font face, elles aussi, à un besoin croissant de modélisation et de gestion des données. La simulation de fluides complexes, de comportement de matériaux dans des conditions extrêmes (forages profonds, transports cryogéniques des gaz, maîtrise de la fusion nucléaire, etc.) et des phénomènes de vieillissement sont des enjeux majeurs de ce secteur. Ces enjeux nécessitent des moyens de simulations puissants qui génèrent de grandes quantités de données, à mettre en lien avec des données expérimentales. Par exemple, le projet VERCORS (VERification Réaliste du CONfinement des RéacteurS) a pour objectif de reproduire, de façon accélérée, le vieillissement du béton constituant les réacteurs nucléaires pour maîtriser et réduire les risques de fuites d'eau irradiée. Ce projet s'appuie sur une maquette réelle, à l'échelle 1/3, équipée de 700 capteurs (jauges de déformations, fibres optiques,...). De nombreuses simulations multi-physiques temporelles doivent être mises en place et corrélées avec l'ensemble des mesures réalisées sur toute la durée des expériences.

Le secteur ferroviaire, quant à lui, privilégie les développements des transports à grande vitesse et la maîtrise de la maintenance de ses équipements. Ces deux axes nécessitent la modélisation et la simulation des systèmes trains/rails pour fiabiliser ce mode de transport tout en augmentant la vitesse et en limitant l'impact sur la durabilité des voies. D'autres solutions sont envisagées, telles que la mise en place de réseaux de capteurs pour une surveillance continue des équipements et une maintenance optimisée.

Le secteur naval est également soumis à des restrictions sur l'impact environnemental. L'économie en carburant et la diminution des émissions de soufre, d'oxydes d'azote et de dioxyde de carbone encouragent les innovations mécaniques par la simulation numérique (allègement des structures, diminution de la traînée) et la recherche en matériaux et traitement de surface (e.g. diminution des frottements). Là encore, la simulation numérique est identifiée comme un facteur clé dans l'amélioration des performances des navires, de même que l'ingénierie collaborative (développement de navires virtuels) et l'échange de données.

De nombreux produits issus des secteurs de l'industrie mécanique (véhicules terrestres, avions, satellites, centrales nucléaires, etc.) sont des produits complexes, intégrant de nombreux composants issus de disciplines diverses et devant satisfaire des exigences fortes et multiples. Ces produits sont développés par tout un ensemble d'entreprises et parties prenantes. L'entreprise étendue représente l'ensemble des parties-prenantes impliquées dans les processus liés au produit, tout au long de son cycle de vie [Le Duigou 10]. Elle concerne les départements, services et équipes internes à l'entreprise, mais aussi ceux des

organisations partenaires, des clients, des sous-traitants, etc.

Les outils numériques pour la simulation et la communication occupent une place importante pour tous ces secteurs pour soutenir leur compétitivité. Par ailleurs, d'autres organismes ont identifiés les grands axes de recherches et de développements pour les années à venir.

Le CEntre Technique des Industries Mécaniques (CETIM) [CETIM 15] identifie les 5 catégories de technologies prioritaires suivantes pour l'industrie mécanique :

1. **Matériaux et Procédés.** Le développement de nouveaux matériaux (composites, élastomères, méta-matériaux, alliages hautes performances, etc.), plus performants et plus adaptés à certains usages, peut offrir par exemple une réduction de la masse du produit ou des frottements, et aboutir à des économies d'énergie. Cette catégorie concerne également l'amélioration des procédés d'assemblage, de fabrication (formage, usinage, etc.), les machines intelligentes et la fabrication additive.
2. **Conception, simulation et essais.** La prise en compte du vieillissement des matériaux, des contraintes d'assemblage et de désassemblage des composants nécessite la considération de l'utilisation et du retrait du produit lors de la phase de conception. L'application des moyens de simulation numérique en amont des phases de conception permet de concevoir des produits et procédés plus fiables, plus sûrs, tout en réduisant le nombre de prototypes nécessaires. La puissance de calcul offerte par les clusters de calcul [High-Performance Computing \(HPC\)](#) permet le développement de simulations précises (multi-échelle et multi-physique) et de moyens de conception collaboratifs (maquette numérique) et immersifs (réalité virtuelle et réalité augmentée).
3. **Mécatronique, contrôle et mesures.** Cette catégorie concerne l'intégration de capteurs (économie d'énergie, communication sans-fil, mesures sans contact), la surveillance continue et le pilotage des procédés et des équipements.
4. **Développement durable.** L'impact environnemental des procédés et des produits est un enjeu majeur de cette décennie, que ce soit par rapport à la consommation énergétique ou par rapport à la génération de déchets nocifs. Cela couvre les démarches d'éco-conception, l'utilisation de matériaux bio-sourcés, la valorisation des déchets ou encore l'amélioration de l'efficacité énergétique des produits et des procédés.
5. **Système d'information et management.** Cela concerne la gestion des moyens informatiques soutenant la compétitivité de l'entreprise. Ces technologies supportent les processus de développements collaboratifs dans un contexte d'entreprise étendue, la gestion des connaissances et des grands volumes de données (Big Data), et la mise en place de maquettes numériques des produits.

Parmi les solutions émergentes pour répondre aux enjeux de l'industrie mécanique, la gestion des données et la simulation numérique sont des solutions transversales. Ces

activités concernent les 5 technologies prioritaires, listées précédemment, identifiées par le CETIM. Les besoins en termes de capacités de gestion, d'échange et d'analyse des données pour la prédiction et l'optimisation sont croissants dans de nombreux secteurs.

L'Institut de Recherche Technologiques (IRT) SystemX, dans sa feuille de route 2016-2020 [IRTSysSystemX 16], se concentre sur quatre programmes de recherches pour répondre aux besoins des entreprises durant les phases de conception de produits : l'ingénierie-systèmes (ingénierie collaborative, produits complexes, entreprise étendue), les transports autonomes (gestion des flux, systèmes embarqués, sécurité), les infrastructures numériques (gestion, échange, analyse et production de données) et les territoires intelligents (logistique urbaine, habitats connectés, optimisation énergétique, sécurité).

La Banque Public d'Investissements (BPI France) a publié en 2015 son second appel à projets portant sur le calcul intensif et la simulation numérique. L'objectif premier de cet appel à projet est de "renforcer la maîtrise des outils permettant d'exploiter les prochaines générations de supercalculateurs et de contribuer au développement de fournisseurs de technologies dans ces secteurs" [BPI 14]. Les secteurs visés sont en priorité ceux de la santé (gestion des données cliniques, diagnostic rapide), de l'urbanisme (gestion temps réels des ressources, simulation des flux) et des industries manufacturières et du secteur de l'énergie (produits complexes, prise en compte des incertitudes, réduction du nombre d'essais, fabrication additive, réduction de modèles). Les axes technologiques considérés portent sur :

1. le développement de nouvelles architectures matérielles et logicielles pour le calcul, le stockage et la gestion des données ;
2. le développement d'outils logiciels, libres et disponibles en ligne, d'analyse et d'interprétation des données (gestion de la parallélisation, réalité virtuelle, visualisation, fouille de données, etc.).

Le développement des capacités informatiques est essentiel pour supporter ces évolutions. L'Union Européenne a lancé en 2014 le programme de recherche Horizon 2020. Ce programme a pour objectif de renforcer la recherche et la compétitivité de l'UE en cohérence avec les enjeux sociétaux actuels, liés à la démographie croissante, le développement durable et les besoins de la mobilité. Un des thèmes de recherches de ce programme concerne le développement de capacités de calculs intensifs **HPC**. Les axes de développements concernent l'architecture matérielle, les algorithmes exploitant ces ressources matérielles et le développement d'un écosystème **HPC** européen durable. L'objectif est de permettre à l'industrie et la recherche de profiter de ressources pour le calcul massivement parallèle et le traitement de grandes quantités de données pour accélérer la recherche et l'innovation dans tous les domaines.

Ces axes de développements identifiés par le CETIM, l'IRT SystemX, BPI France et

l'UE, concernent eux aussi le développement des outils numériques au sein des activités de la mécanique. Les activités de simulation numérique et de gestion de données sont ainsi identifiées par plusieurs organismes comme étant des axes clés de développement dans les années à venir.

Tous ces enjeux, tant industriels que scientifiques, relevés par les acteurs de l'industrie mécanique, forment le contexte de ces travaux de recherche. Ceux-ci se focalisent plus particulièrement sur les problématiques de simulation numérique et de gestion de données au cours du processus de conception de produit. Ces deux domaines occupent une place importante pour assurer la compétitivité des entreprises. Les sous-sections suivantes détaillent respectivement la place de la gestion du cycle de vie du produit et de la simulation numérique dans le contexte industriel pour en relever les verrous scientifiques.

### 1.1.2 Gestion du cycle de vie du produit

L'industrie mécanique est confrontée au développement de produits complexes, à travers des entreprises étendues, et à la nécessité de rester compétitive sur les différents marchés économiques. De nombreux verrous industriels concernent l'amélioration du produit, la réduction de son coût, la rapidité de sa mise sur le marché, les processus d'ingénierie collaborative et la prise en compte de tout le cycle de vie du produit.

Le cycle de vie du produit peut être considéré selon les deux aspects suivants.

- Le cycle de vie du produit réel (Figure 1.2), qui prend en compte le produit comme une entité physique existante. Cette vision du cycle de vie débute par l'acquisition de la matière brute et transformation, se poursuit par la fabrication du produit, son utilisation et son retrait pour destruction, stockage ou recyclage.
- Le cycle de vie du produit virtuel (Figure 1.3), qui considère le produit sous forme numérique et idéalisé. Ce cycle débute par la phase de conceptualisation, qui se poursuit par les phases de conception préliminaire et de conception détaillée. Le reste de ce cycle rejoint le cycle de vie du produit physique pour apporter un support virtuel aux phases de fabrication, d'utilisation et de retrait [Le Duigou 10, IFIP/IFAC 99].

Une autre description [Terzi 10] décompose le cycle de vie par trois phases principales. Le début de vie (**Beginning Of Life (BOL)**) comprend les phases de conception et de fabrication. Le cycle se poursuit par la distribution, l'utilisation et la maintenance du produit (**Middle Of Life (MOL)**). La fin de vie concerne le retrait du produit, son recyclage, son désassemblage (**End Of Life (EOL)**).

La gestion du cycle de vie du produit (**Product Lifecycle Management (PLM)**) est une stratégie d'entreprise qui a pour objectif principal d'améliorer la maîtrise du produit tout au long de son cycle de vie [Stark 04]. La stratégie **PLM** vise à supporter les processus



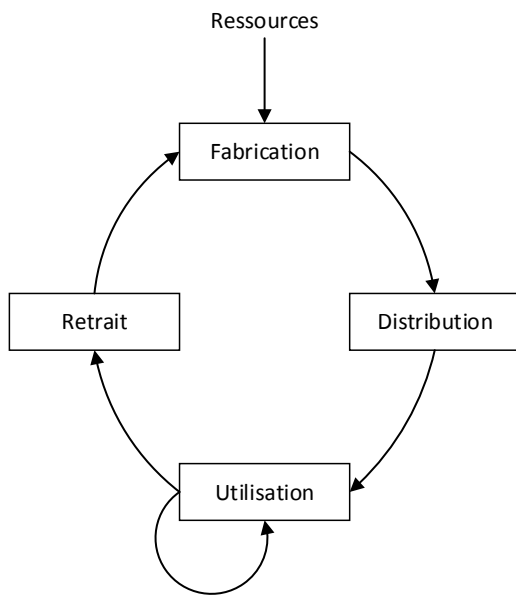


FIGURE 1.2 – Cycle de vie du produit réel, adaptée de [Le Duigou 10]

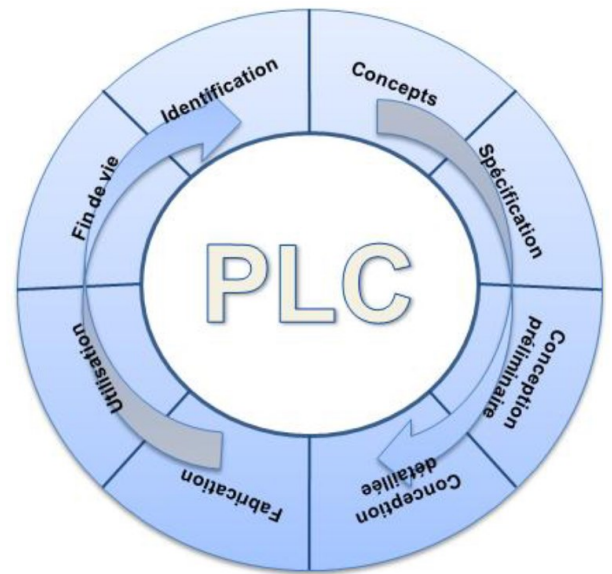


FIGURE 1.3 – Cycle de vie du produit virtuel, tirée de [Le Duigou 10]

collaboratifs liés au produit au sein d'entreprises étendues, ainsi que tous les éléments de définition du produit, tout au long du cycle de vie du produit [CIMDATA 03]. Le **PLM** est une approche centrée sur le produit, impliquant l'utilisation de nombreux outils informatiques pour la capitalisation, la gestion, le partage et la réutilisation des données et des connaissances en entreprises, dans le but d'atteindre les objectifs de qualité, de coût et de délai [Terzi 10].

L'application du **PLM** repose notamment sur des outils de gestion des données, utilisées et produites durant le cycle de vie et par tous les acteurs et tous les processus. Plusieurs types de systèmes existent, correspondant à certaines phases spécifiques du cycle de vie du produit. Les systèmes **Product Data Management (PDM)** gèrent principalement les données durant la phase de conception, et plus spécialement les données portant sur l'architecture du produit, la nomenclature, et la définition géométrique. Ces systèmes sont à l'origine de la stratégie **PLM**. Les systèmes de gestion de données de simulation sont plus récents, et concernent spécifiquement les données et les processus de simulations numériques menés lors des phases de conception (**Simulation Data Management (SDM)**) [CIMDATA 12]. Ces aspects seront particulièrement détaillés dans la suite de ce chapitre. D'autres systèmes sont utilisés dans l'industrie, tels que les logiciels de gestion des relations clients (CRM - Customer Relationship Management), les systèmes gérant la phase de fabrication (MPM - Manufacturing Process Management - MES - Manufacturing Execution System - et TMS - Tool Management System), les systèmes gérant les données liées à la logistique (SCM - Supply Chain Management) et des solutions plus générales à l'échelle de l'entreprise (ERP - Enterprise Resource Planning).

Par exemple, les fonctionnalités principales des systèmes **PDM** sont les suivantes [CIM-DATA 03] :

- la capitalisation des données de définition du produit (nomenclature, données expérimentales, etc.), des décisions prises sur l'ensemble du processus de développement ;
- l'accessibilité et le partage sécurisé de ces données. Les données pertinentes doivent être accessibles aux personnes concernées et aux moments opportuns. Cela implique une gestion des droits d'accès et la constitution d'une vue du produit adaptée à l'utilisateur ;
- la traçabilité et le maintien de l'intégrité des données de définition tout au long du cycle de vie. Cela inclut la gestion des changements, des versions et des configurations ;
- la gestion et le maintien des processus liés aux produits, tels que les processus de conception, de fabrication, d'expérimentation, etc.

Le **PLM** tend à fédérer tous les acteurs du cycle de vie du produit et tous les outils nécessaires autour d'une description unifiée du produit. Les processus collaboratifs sont alors favorisés et permettent le développement de produits complexes, multi-disciplinaires et soumis à des exigences nombreuses et fortes. L'application d'une stratégie **PLM** est confrontée à plusieurs verrous scientifiques, tels que [CIMDATA 16]<sup>1</sup> :

- l'extension des applications aux phases **MOL** et **EOL**. La maîtrise du produit au-delà de sa conception et de sa fabrication permet, par exemple, d'améliorer le produit par rapport aux conditions d'utilisation réelles, d'optimiser les opérations de maintenances ou encore d'améliorer la réutilisabilité du produit [Terzi 10]. Une tendance majeure est le close-loop **PLM**. Des retours d'informations continues sont possibles tout au long du cycle de vie du produit réel vers les phases amonts [Jun 07], par l'intégration de capteurs et de dispositifs d'identification au sein mêmes des produits. Chaque évènement concernant le produit est alors capitalisé, ce qui conduit à l'amélioration de la traçabilité du produit. Chaque acteur du cycle de vie du produit peut alors réutiliser une partie de ces données pour améliorer les prises de décisions. Les développements récents se tournent vers le concept de "digital twin" [Grieves 14]. Le produit réel et le produit virtuel sont connectés par le biais de capteurs intégrés au produit réel et communiquant les informations au produit virtuel durant tout son cycle de vie pour, par exemple, effectuer des prédictions du comportement du produit en temps réel par rapport aux données collectées [Patterson 16]. Cet axe de développement est lié à la problématique du "Big Data", s'attachant à gérer des quantités massives de données fortement hétérogènes, à les rendre disponibles et à les interpréter rapidement [Siddiqi 16, Fahad 14] ;

---

1. <http://www.cimdata.com/en/events/plm-conferences/event/175-plm-road-map-2016-for-the-automotive-industry>

- l'intégration des aspects environnementaux et l'éco-conception dans la stratégie **PLM** [Vila 15]. Cette évolution vise le développement de produits impliquant un impact environnemental minimal au niveau de la fabrication (procédés et matériaux), des déchets produits en cours d'utilisation et des opérations de maintenance, et du recyclage du produit. L'impact environnemental pourrait alors être maîtrisé sur tout le cycle de vie du produit. Par exemple, l'empreinte carbone du produit pourrait être prédite et minimisée dès la phase de conception ;
- la collaboration de multiples disciplines autour d'un même produit, qui produit des problèmes d'interopérabilité entre les différents outils. L'utilisation de standards pour l'échange de données et d'ontologies pour le partage des connaissances sont utilisés pour soutenir les processus collaboratifs [Bruno 15, Haas 16, Draghici 09, CRESCENDO 13, Pardessus 00]. Une évolution notable est celle de la maquette numérique. Cette maquette rend accessible les données décrivant le produit virtuel (géométrie, comportement, etc.) à toutes les parties prenantes. Cette technologie facilite les démarches de conception collaborative dans une entreprise étendue ;
- l'intégration plus en amont des activités de simulations numériques, pour simuler le comportement du produit et des processus liés, tout au long du cycle de vie ;
- la prise en compte de nouveaux procédés de fabrication, tels que la fabrication additive et l'utilisation de composites et de méta-matériaux.

La stratégie **PLM** offre aux entreprises des moyens de relever les défis actuels de l'industrie mécanique. La gestion et la capitalisation des données restent essentielles pour maîtriser le produit, l'améliorer et capitaliser les connaissances de l'entreprise. Parmi les processus de développement des produits, la simulation numérique apporte des solutions puissantes aux entreprises pour conserver leur compétitivité, mais peut exiger des connaissances expertes avancées et peut générer de grandes quantités de données. Ces aspects sont détaillés dans la suite de ce chapitre.

### 1.1.3 Simulation numérique

Les activités de simulations numériques ont pris une place centrale pour le développement de produits. La simulation numérique est identifiée par de nombreux secteurs de la mécanique française comme un domaine clé de la recherche et du développement pour garantir leur compétitivité future.

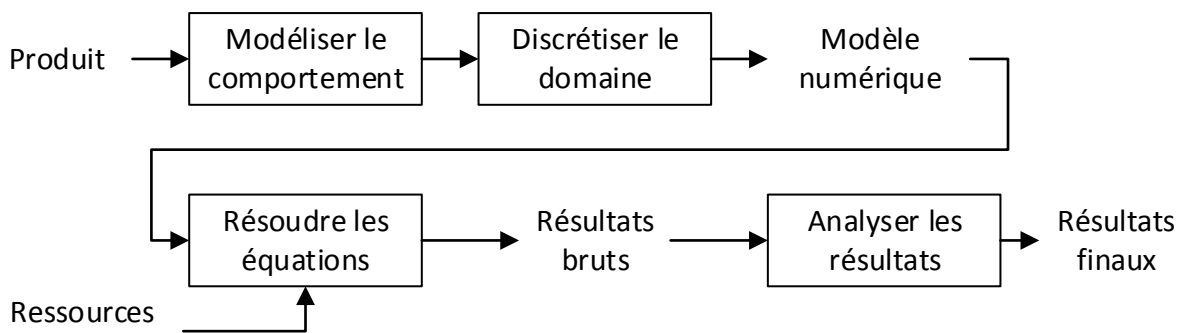
Les activités de simulation dans l'industrie consistent à reproduire le comportement réel d'un système en le modélisant aussi précisément que nécessaire et aussi simplement que possible, pour en prédire le comportement et, si besoin, l'optimiser. La simulation numérique est fortement liée aux disciplines expérimentales, nécessaires pour valider un

modèle de simulation. Dans le cadre des développements de produits mécaniques, les activités de simulations peuvent concerner toutes les phases du cycle de vie, telles que les phases de conception, de fabrication, de distribution ou d'identification des défaillances.

La simulation numérique consiste à mettre la puissance de calcul informatique au service des activités de simulation. Cette évolution, suivant les progrès des technologies de l'information, a permis d'obtenir des prédictions de comportement plus précises, plus rapidement, et pour des phénomènes et des systèmes plus complexes. Pourtant, toutes ces possibilités peuvent provoquer un besoin d'expertises avancées et la génération d'une grande quantité de données hétérogènes.

Le processus de simulation numérique peut être décomposé en plusieurs étapes (Figure 1.4) [Adams 08, Pourroy 99].

FIGURE 1.4 – Le processus de simulation numérique



L'étape de modélisation consiste à créer une représentation simplifiée, mais suffisamment précise, du(des) phénomène(s) étudié(s), du produit et de son environnement [Maria 97] :

- de nombreux phénomènes peuvent être simulés, tels que les phénomènes mécaniques, thermiques, électromagnétiques et chimiques. Ces phénomènes peuvent dépendre du temps (phénomènes dynamiques), être non-linéaires, instables et couplés entre eux. Le comportement du produit, soumis à un phénomène donné, est décrit à l'aide de lois de comportement ;
- la représentation géométrique du produit est réalisée à l'aide d'outils de **Conception Assistée par Ordinateur (CAO)**. Cette représentation idéale du produit peut-être encore simplifiée. Les éléments géométriques n'apportant pas de précision supplémentaire aux résultats souhaités sont simplifiés ou supprimés, pour préparer l'étape suivante de discrétisation et pour diminuer le coût et le temps de calcul ;
- l'environnement du produit comprend les sollicitations externes, les conditions aux limites et les conditions initiales, lorsqu'il s'agit d'un problème temporel.

L'étape de discrétisation rend la simulation exécutable par un ordinateur. Les lois continues, décrivant le comportement du système étudié, sont transformées en équations

discrètes. Le problème global est partitionné en un nombre fini de sous-domaines élémentaires simples (maillage), en espace et en temps. Cette étape définit le nombre d'inconnues à calculer (degrés de liberté) et conditionne directement le coût de l'exécution et la quantité de données générées. Il existe plusieurs méthodes de discrétisation, telles que la méthode des différences finies, des éléments finis, des éléments de frontière ou encore des volumes finis.

A l'issue des étapes de modélisation et de discrétisation, le modèle numérique est obtenu et est prêt à être exécuté pour estimer les réponses du produit. Cette étape est celle exigeant le plus de puissance de calcul. Elle peut nécessiter plusieurs itérations, dans le cas de problèmes non-linéaires, de problèmes temporelles ou de problèmes d'optimisation. De même, cette étape peut être constituée de plusieurs simulations différentes chaînées, pour des simulations multi-physiques, multi-échelles. Le temps d'exécution d'une simulation doit être maîtrisé pour assurer un processus de simulation efficace. Il dépend notamment de la complexité de l'algorithme de résolution, du nombre de degrés de liberté du problème et de la puissance de calcul disponible et exploitable par le programme de résolution. Les ressources matérielles nécessaires pour l'exécution dépendent de la complexité des modèles. Cela peut s'échelonner des calculs locaux, sur un ordinateur personnel, aux calculs sur des clusters dédiés, type [HPC](#). Concernant cet aspect, il est important de préciser la distinction entre le coût et le temps d'exécution. Les sources de coûts d'exécutions proviennent de l'acquisition de la puissance de calcul nécessaire, le coût des capacités logicielles (licence) et le coût d'immobilisation de ces ressources (ralentissement ou arrêt des études des collaborateurs). La puissance de calcul peut être délivrée par une machine locale, un cluster interne ou par les offres de cloud-computing. Le temps d'exécution tend à diminuer en fonction de la puissance allouée, ce qui peut, selon le type d'offre de puissance de calcul, faire augmenter plus ou moins rapidement le coût du calcul.

Enfin, les résultats bruts sont analysés lors de la phase de post-traitement, pour déterminer si le modèle est fonctionnel et s'il représente correctement la réalité. Si le modèle s'avère être conforme aux résultats expérimentaux, les résultats auront alors un sens physique et pourront être comparés aux exigences du client. Cette étape permet d'estimer les erreurs accumulées au cours du processus [Benhafid 05], provenant de l'accumulation des erreurs théoriques, des erreurs d'idéalisation, de discrétisation, de troncature et d'arrondis.

En termes d'axes de développements, la NASA [NASA 15] a identifié principalement les points suivants :

- les capacités de calcul, dont le développement de supercalculateurs "exascale" et d'ordinateurs quantiques, l'intégration des problématiques liées au "Big Data";
- la modélisation, et plus particulièrement l'utilisation des modèles numériques comme outils centraux durant tout le cycle de vie du produit, la modélisation de l'être

humain en interaction directe avec le produit, ou encore, l'amélioration des outils pour intégrer les aspects collaboratifs et l'échange de données, et la réutilisation de modèles ;

- la simulation. Cet axe comprend le calcul distribué (sur des sites géographiquement éloignés), l'utilisation d'outils **SDM**, l'amélioration de la précision des simulations pour développer un clone numérique du produit ("digital twin"), le recours aux simulations multi-physiques (pour les phénomènes complexes), multi-fidélités (pour les différentes parties-prenantes) et multi-échelles (de l'atome à la structure), la prise en compte des incertitudes et des processus aléatoires et stochastiques ;
- la gestion des données. Cela couvre la traçabilité des données de simulation sur tout le cycle de vie, la capacité d'analyser ces données automatiquement pour supporter les prises de décisions ou encore la sécurisation des données.

Le processus de simulation numérique est un processus complexe. De nombreux choix conditionnent la qualité des résultats finaux. Ce processus peut nécessiter un temps important pour modéliser le problème, le discrétiser, le résoudre et analyser ses résultats. Il peut être nécessaire de réaliser plusieurs essais avant de trouver la bonne stratégie de modélisation/discrétisation/exécution.

Dans la suite de ces travaux, le modèle numérique produit par ce processus est supposé vérifié, validé et totalement opérationnel. L'utilisation de ce modèle numérique peut être étendue par la prise en compte des variabilités et des incertitudes liées au produit.

#### 1.1.4 Prise en compte des variabilités

La prise en compte des incertitudes et des variabilités liées au produit est une étape clé pour assurer la qualité d'un produit. Elle impacte tous les secteurs de l'industrie mécanique. L'étude des variabilités des propriétés, des performances et des conditions d'utilisation d'un produit permet d'estimer son comportement dans une modélisation plus réaliste. Les caractéristiques géométriques, les propriétés des matériaux, les sollicitations externes ne sont plus constantes au sein d'un modèle numérique idéal. La fiabilité - capacité d'un système à accomplir les fonctions requises dans des conditions fixées pendant une durée donnée [Arlat 96] - , la robustesse - capacité d'un système à conserver ses performances lorsque les conditions d'utilisation changent [Taguchi 00] - et la sensibilité - influences de la variabilité des paramètres du système sur ses performances [Saltelli 02]- du produit peuvent alors être étudiées dès la phase de conception.

Parmi les verrous scientifiques identifiés concernant la prise en compte des incertitudes pour la mécanique [AFM 15], on peut citer les besoins :

- d'améliorer la compréhension entre le domaine des essais physiques et le domaine de la simulation numérique ;
- d'améliorer l'applicabilité des méthodologies de représentation de l'incertain. L'application de ces méthodologies est puissante mais leurs applications restent laborieuses pour des cas industriels ;
- de propager les incertitudes au sein des simulations multi-échelles et multi-physiques ;
- de représenter ces incertitudes au sein d'une vision globale du produit tout au long de son cycle de vie.

Une approche possible pour rendre compte des variabilités en phase de conception est l'utilisation d'une démarche de **PEN**. Cette démarche, décrite dans la section suivante, consiste à définir une série de simulations en faisant varier les paramètres du modèle numérique. Cela permet de renforcer la fiabilité et la robustesse d'un produit en amont du processus de développement. Les expériences sont alors effectuées numériquement et conduisent à une réduction du nombre de prototypes réels et du nombre d'essais réels. De plus, les résultats d'un **PEN** peuvent être comparés avec les résultats d'un plan d'expériences réelles afin d'améliorer le modèle numérique.

Là où la simulation numérique peut exiger une expertise avancée et peut générer une grande quantité de données, la démarche des **PEN** vient multiplier la quantité de données générées, le coût et/ou le temps d'exécution des calculs (par le nombre d'expériences) et amener d'autres méthodes nécessitant des connaissances expertes supplémentaires. Par exemple, pour une simulation nécessitant 1 heure de calcul et générant 1Go de données résultantes, un **PEN** comprenant 100 expériences nécessitera plus 4 jours de calculs intensifs avant d'obtenir et de pouvoir analyser les 100Go de données générées. Ceci représente un frein pour la prise de décision et la réactivité des entreprises.

Ces travaux de recherche ont été menés dans le cadre du projet FUI 16 SDM4DOE (Simulation Data Management for Design of Experiments), co-labellisé par les pôles de compétitivité ID4CAR<sup>2</sup> et SYSTEMATIC<sup>3</sup>. Le consortium de ce projet est constitué :

- d'un concepteur et fabricant d'équipements pour les constructeurs automobiles, Valeo (Systemes Thermiques SAS branche Thermique Habitable)<sup>4</sup> ;
- d'une société d'ingénierie avancée en mécanique et génie civil, NECS<sup>5</sup> ;
- d'une société d'expertise en ingénierie et en informatique scientifique spécialisé dans le calcul des structures, Alneos<sup>6</sup> ;

---

2. <http://www.id4car.org/>

3. <http://www.systematic-paris-region.org/>

4. <http://www.valeo.com/fr/>

5. <http://necs.fr/>

6. <http://www.alneos.com/>

- du Laboratoire de Génie Logiciel pour la Simulation (LGLS) du Commissariat à l'Énergie Atomique et aux énergies alternatives (CEA) <sup>7</sup> ;
- du laboratoire de mécanique Roberval de l'Université de Technologie de Compiègne (UMR UTC/CNRS 3773) <sup>8</sup> ;
- du Laboratoire des Sciences du Numérique de Nantes (LS2N, UMR6004) <sup>9</sup> .

Ce projet répond à des problématiques de gestion, de capitalisation et de traçabilité des données générées par un **PEN**, ainsi qu'à des problématiques de réduction de temps d'exécution et d'accélération des prises de décisions par rapport aux résultats des **PEN**. Le résultat est un logiciel open-source capable de gérer et d'exécuter des processus de **PEN**, tout en mettant en œuvre des méthodes avancées de traitement des résultats des **PEN** et des ressources de calculs de type serveur **HPC**.

SDM4DOE allie les enjeux de la gestion de données dans un contexte collaboratif et les enjeux de prise en compte des incertitudes et des variabilités pour proposer une solution générique et exploitable par tous les acteurs du domaine de la mécanique, dans le but d'améliorer leurs produits. La fiabilité et la robustesse du produit pourront être améliorées tout en réduisant la durée du développement des produits. Les deux sections suivantes détaillent ces aspects pour en dégager la problématique des recherches présentées ici.

---

7. <http://www-centre-saclay.cea.fr/>

8. <http://roberval.utc.fr/>

9. <https://ls2n.fr/>



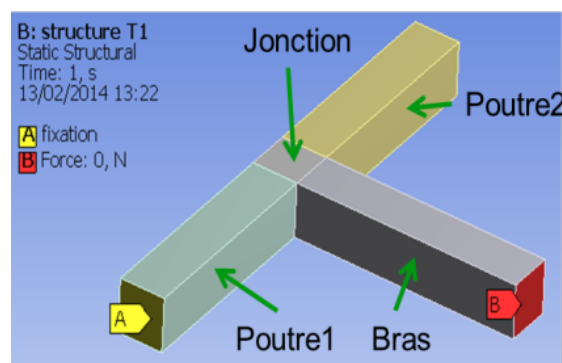
## 1.2 Processus de plan d'expériences numériques

Les performances finales et les propriétés d'un produit sont rarement identiques d'un exemplaire à un autre, et peuvent évoluer au fil du temps. Ces variabilités peuvent être caractérisées et prises en compte en réalisant plusieurs essais, pour améliorer la qualité du produit, et notamment sa robustesse et sa fiabilité. L'approche des plans d'expériences permet de prendre en compte ces variabilités en réalisant plusieurs expériences avec des valeurs de paramètres différentes, tout en limitant le nombre d'expériences nécessaires.

Le processus de PEN réalise les expériences par le biais du modèle numérique. Il fournit des informations pour une meilleure compréhension du comportement du produit, une optimisation de ses performances et pour des prises de décisions rapides en amont du processus de conception. Il est alors possible, par exemple, d'évaluer virtuellement le comportement du produit sur toutes les plages de variations de ses dimensions tolérancées. Ce type d'application conduit à une réduction du nombre d'essais physiques nécessaires et à un processus de développement plus court, dû à des prises de décisions plus précoces.

Cette section définit le processus de PEN et détaille ses différentes étapes. Les différentes méthodes de planifications des expériences et d'analyse des résultats de ces expériences sont détaillées. Une étude simplifiée est utilisée pour illustrer ces concepts. Cette étude, utilisée durant le projet SDM4DOE, porte sur une structure mécanique théorique et simple (Figure 1.5). Les caractéristiques principales de ce modèle, nommé T1, sont résumées dans le Tableau 1.1.

FIGURE 1.5 – Représentation 3D de la structure du modèle T1



L'application des plans d'expériences à la simulation numérique est soumise aux spécificités du processus de simulation numérique. Trois spécificités majeures peuvent être relevées, par rapport aux plans d'expériences réelles :

1. un modèle numérique peut comporter un très grand nombre de facteurs, avec de larges intervalles de variations et de multiples étapes de calculs (i.e. simulations multi-physique). Le comportement du système est alors difficile à analyser et requiert l'exécution d'un très grand nombre d'expériences ;

TABLEAU 1.1 – Caractéristiques principales du modèle T1

Nom du modèle	Type de simulation	Paramètres			Réponses	
T1	Statique linéaire - 56 éléments finis linéaires hexahédriques	E	F <sub>x</sub>	F <sub>y</sub>	Reac	Strain

Légende : **E** : module d'Young (matériau homogène, isotrope, élastique linéaire), loi uniforme entre 0.5 et 2 MPa; **F<sub>x</sub>** : force horizontale (x) en B, loi uniforme entre 2N et 10N; **F<sub>y</sub>** : force verticale (y) en B, loi uniforme entre 2N et 10N; **Reac** : force résultante aux encastremets; **Strain** : maximum du premier invariant du tenseur des déformations.

2. la réponse étudiée peut-être fortement non-linéaire, non-monotone, surtout lorsqu'elle résulte de nombreuses simulations couplées ou chaînées;
3. dans la plupart des cas, les simulations numériques sont suffisamment déterministes pour ne pas nécessiter de répétitions d'expériences. Toutefois, il existe trois situations pour lesquels il est nécessaire de répéter plusieurs fois une même expérience :
  - la chaîne de calcul dispose d'un remaillage automatique. Une expérience de contrôle peut être nécessaire pour quantifier la variance de la réponse par rapport au changement de maillage;
  - certains codes de calculs peuvent être volontairement probabilistes pour reproduire une incertitude expérimentale;
  - l'erreur de résolution ne peut plus être négligée et doit être prise en compte comme incertitude expérimentale.

### 1.2.1 Définition d'un plan d'expériences

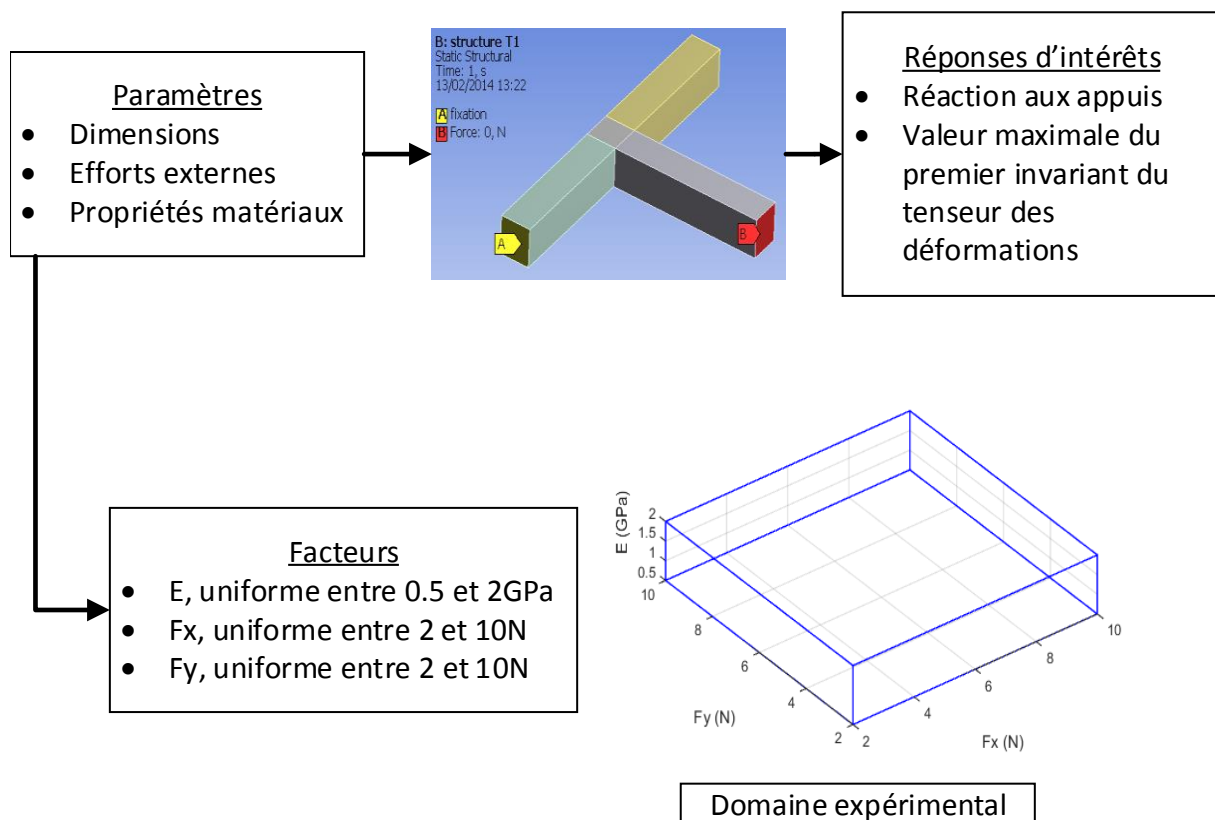
Les plans d'expériences ont d'abord été utilisés dans un contexte d'expériences réelles. La définition des plans d'expériences, numériques ou non, reposent sur les quatre étapes suivantes, selon l'objectif et les contraintes de la démarche (sous-section 1.2.3) :

1. la définition du système, qui est représenté par un comportement, s'exprime par des réponses, et dépend de variables d'entrées. L'ensemble des variables d'entrée est composé des paramètres contrôlables du système et des variables incontrôlables, mesurables ou non (ex : la température extérieure);
2. le choix des réponses d'intérêt, parmi les réponses mesurables et selon les exigences associées au produit;
3. le choix et la définition des facteurs du plan d'expériences. L'ensemble des facteurs appartient à l'ensemble des paramètres contrôlables du modèle numérique initial. Le choix des facteurs est réalisé selon les connaissances disponibles quant à la causalité entre les facteurs et la réponse d'intérêt préalablement choisie. On associe à chaque facteur un domaine de variation, continu ou discret, qualitatif ou quantitatif. Le

domaine résultant de l'intersection de tous les domaines de variation constitue le domaine expérimental (Figure 1.6). Chaque facteur dispose d'un nombre fini de niveaux ou de modalités, qui appartiennent au domaine de variation ;

4. la méthode de planification des expériences. Un type de plan est choisi pour définir une répartition optimale des expériences dans le domaine expérimental, selon des critères et des contraintes. Cette étape est détaillée dans la sous-section suivante.

FIGURE 1.6 – Définition d'un plan d'expériences numériques pour le modèle T1.



À chaque expérience correspond, après son exécution, une réponse particulière du système. Un modèle mathématique empirique est associé au plan d'expériences pour analyser les réponses. Cela permet, par exemple, d'obtenir une surface de réponses continue à partir des  $n$  réponses (scalaires) du système obtenues.

Un plan d'expérience constitue un moyen d'échantillonner le domaine expérimental dans un but précis. L'utilisation d'un plan d'expériences, et, par extension, d'un PEN, s'inscrit dans un processus à part entière, défini par un objectif, des contraintes et des méthodes pour atteindre cet objectif et respecter ces contraintes. Avant de passer en revue les différentes étapes d'un processus de PEN, les principaux objectifs des PEN sont décrits dans la sous-section suivante.

### 1.2.2 Objectifs des plans d'expériences numériques

Les PEN sont utilisés principalement pour quatre objectifs distincts :

- l'analyse de sensibilité classe les facteurs du PEN, ainsi que leurs interactions, suivant leurs influences sur la réponse d'intérêt. Cela permet d'identifier les facteurs et les interactions réellement influents et ceux qui peuvent être négligés. L'analyse de sensibilité limite ainsi le nombre d'expériences, la mauvaise répartition des expériences dans le domaine expérimental, le coût de calcul des méthodes d'analyse des résultats, etc. [Iooss 14] ;
- la production d'un modèle de substitution (appelé aussi méta-modèle) consiste à produire un modèle empirique, à partir du PEN et de ses résultats, capable de reproduire fidèlement le comportement du modèle numérique initial, en tout point du domaine expérimental. Le modèle de substitution se substitue au modèle numérique initial dans tout le domaine expérimental, tout en ayant un coût d'exécution plus faible. Ainsi, un modèle de substitution est un outil permettant d'accélérer les processus de conception et de prise de décision ;
- l'analyse de robustesse s'intéresse à la tenue des exigences du produit en prenant en compte les variabilités sur ses paramètres (dimensions, matériaux, chargements, etc.). Les PEN permettent de générer un ensemble d'expériences suivant des lois de probabilité données. Il est alors possible de simuler ces variabilités et d'obtenir le comportement du produit résultant. L'analyse de robustesse vise à minimiser l'influence de la variabilité des paramètres du système sur ses performances [Fowlkes 95] ;
- l'optimisation du produit consiste à modifier ses propriétés pour en maximiser ses performances sous certaines contraintes. Il peut s'agir par exemple de minimiser la masse du produit tout en conservant un certain nombre de surfaces fonctionnelles et un niveau de contraintes mécaniques limité.

### 1.2.3 Définition d'un processus de plan d'expériences numériques

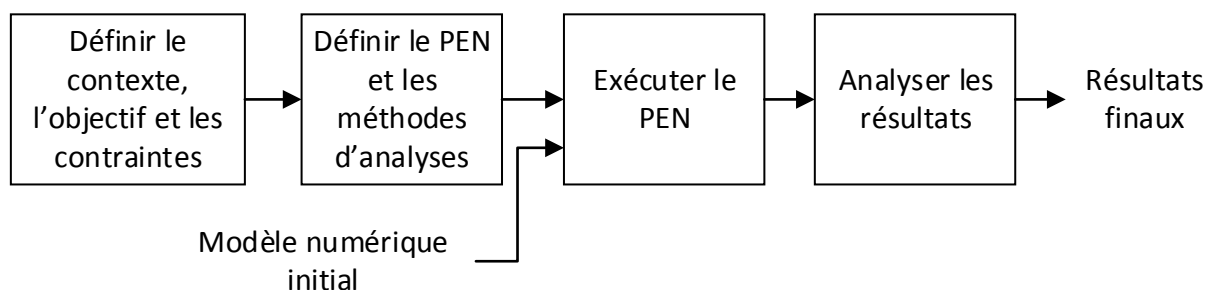
Un PEN s'inscrit au sein d'un processus mobilisant différents outils numériques pour atteindre un objectif fixé. Dans un objectif général de réduction des coûts, la démarche de plan d'expériences peut être réalisée en plusieurs processus de PEN à coût d'exécution croissant, jusqu'à obtenir le résultat désiré. Par exemple, dans le cas d'un modèle numérique initial présentant un coût d'exécution très élevé et de nombreux facteurs, il convient de réaliser d'abord une analyse de sensibilité. Si le nombre de facteurs est trop important, plusieurs analyses de sensibilité peuvent être menées successivement [Iooss 14]. Une fois le nombre de facteurs réduit, un processus de PEN de production d'un modèle de substitution peut être entrepris. Là encore, plusieurs processus de PEN, de coût croissant, peuvent être

mis en place successivement. Une fois le modèle de substitution obtenu, il peut être utilisé pour une analyse de sensibilité plus précise ou pour un objectif d'optimisation.

Un processus de **PEN** est constitué de quatre étapes principales, qui sont (Figure 1.7) :

1. la définition du contexte, de l'objectif et des contraintes. Cela comprend les données définissant le produit étudié, le projet correspondant, le modèle numérique initial ainsi que toutes les données assurant la traçabilité du processus de **PEN** et sa capitalisation. L'objectif est choisi parmi ceux relevés dans la sous-section 1.2.2. Les contraintes peuvent concerner une limite de temps d'exécution, des limitations techniques de réalisation des expériences ou sur un budget expérimental limité (chaque expérience à un coût et une durée) ;
2. la définition du **PEN** (sous-section 1.2.1), ainsi que de toutes les méthodes utilisées pour atteindre l'objectif fixé, sous les contraintes définies. Il peut s'agir d'une méthode d'analyse de sensibilité ou encore d'une méthode de production d'un modèle de substitution. Ces différents choix sont détaillés dans les sous-sections suivantes. L'ensemble de ces choix constitue la configuration du processus de **PEN** ;
3. l'exécution des expériences. Chaque expérience est communiquée au logiciel de simulation pour obtenir la réponse correspondante du modèle. La réponse est sauvegardée pour une analyse ultérieure ;
4. l'analyse des résultats du **PEN** et l'obtention des résultats finaux. Les résultats bruts sont analysés suivant l'objectif fixé. Les méthodes d'analyse de sensibilité, de production d'un modèle de substitution, de robustesse ou d'optimisation sont appliquées. Les résultats du **PEN** et des méthodes d'analyses sont vérifiés et synthétisés.

FIGURE 1.7 – Le processus de PEN



L'utilisateur du processus de **PEN** doit définir ainsi tous les éléments nécessaires pour exécuter le **PEN** et les méthodes d'analyse des résultats pour atteindre l'objectif tout en respectant les contraintes du processus. Pour chacun de ces choix, de nombreuses possibilités s'offrent à l'utilisateur, rendant la tâche complexe.

Les sous-sections suivantes détaillent les choix à réaliser afin d'identifier plus précisément les problèmes qui se poseront aux utilisateurs.

### 1.2.4 Types de plan d'expériences

Un plan d'expériences définit un ensemble d'expériences réparties dans le domaine expérimental. Il existe de nombreux types de plans d'expériences et le choix de l'un d'eux peut s'avérer difficile. Le principe de ce choix est de sélectionner le plan le plus rentable, présentant le moins d'expériences possible pour obtenir le plus d'informations possible. Ce choix dépend de l'objectif fixé, des méthodes d'analyse utilisées, des contraintes expérimentales, des ressources, du comportement étudié (lié aux facteurs et aux variables d'intérêt) et de l'expertise à disposition. Cette section dresse une liste non-exhaustive et synthétique des plans d'expériences existants.

Deux familles de types de plans peuvent être distinguées. D'une part, les plans "usuels" ont été conçus pour des applications sur des systèmes réels, pour lesquels les expériences peuvent être très coûteuses, longues et difficiles à mettre en place. D'autre part, les types plus spécifiquement utilisés pour des applications numériques, dont le but principal est de générer une distribution d'expériences remplissant de la meilleure façon le domaine expérimental.

#### Plans usuels

Les plans usuels sont associés à un modèle empirique polynômial pour analyser les résultats. Parmi ces plans, les plans factoriels complets sont les plus simples à concevoir mais sont très coûteux, puisqu'ils couvrent toutes les combinaisons de niveaux entre tous les facteurs. Il est moins coûteux d'utiliser des plans factoriels fractionnaires. Ces plans sont des sous-ensembles de plans factoriels complets, ce qui conduit à négliger, *a priori*, l'influence de certains facteurs ou interactions sur la réponse [Montgomery 00]. D'autres plans très courants proposent un nombre limité d'expériences pour la modélisation de surface de réponse (modèle empirique) et l'analyse de sensibilité, tel que les plans composites, les plans de Box-Behnken, les plans de Doehlert, les plans de Taguchi et les plans de Plackett-Burmann [Hedayat 78, Khuri 10]. Certains de ces plans sont aussi limités en termes de nombre de facteurs. Les plans de type produit (ou croisés) [Taguchi 00] sont utilisés plus spécifiquement pour l'analyse de robustesse. Ils combinent un plan à partir des facteurs de contrôles et un plan à partir des facteurs de bruit connus et relativement maîtrisables ou quantifiables (ex : conditions de température et de pression ambiante) en phase d'expérimentation. Les plans de mélanges [Goupy 00] intègrent une contrainte sur les niveaux des facteurs (ex : étude de concentration de différents constituants). Les réseaux de Scheffé sont couramment utilisés pour traiter ce type de cas.

Il existe une famille de plans optimaux conçus pour produire un modèle empirique de qualité, basés sur la minimisation de la variance du modèle empirique [Gazut 08]. Les

critères d’optimalités, tels que les critères D, A, E, G et O, définissent les expériences de manière à minimiser une mesure de cette variance. D’autres critères d’optimalité existent, comme la J-optimalité, qui minimise la variance et le biais du modèle estimé [Box 59]. Ces critères peuvent définir des types de plans par le moyen d’un algorithme d’optimisation, mais peuvent aussi caractériser un plan d’un autre type. Par exemple, un plan factoriel complet à 2 niveaux est O-optimal par définition.

Ces plans d’expériences usuels ne sont pas toujours adaptés au contexte numérique mais restent utilisables. Certains intègrent des répétitions d’expériences, ce qui est inutile si l’on considère un modèle numérique déterministe et répétable. De plus, la plupart sont limités à un nombre donné de niveaux par facteurs et à un type de modèle empirique imposé. Ils n’offrent pas toujours un échantillonnage uniforme dans l’espace de conception.

### Plans ”Space-Filling”

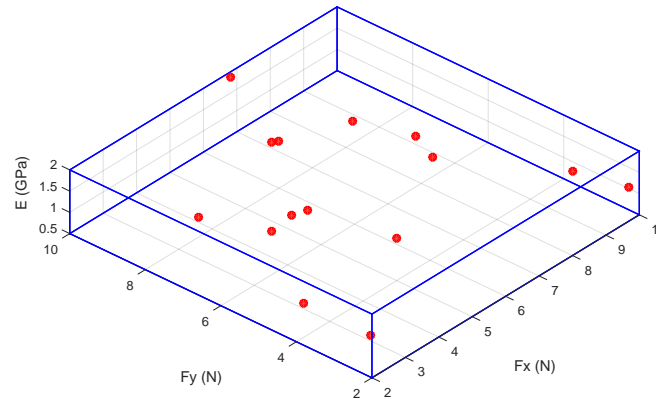
Les plans d’expériences ”Space-Filling” (remplissage de l’espace) sont plus spécifiques aux applications numériques. Ils permettent de définir un grand nombre d’expériences et de traiter un grand nombre de facteurs et de niveaux. De nombreux plans sont basés sur la recherche d’une distribution des expériences la plus uniforme possible, en évitant les amas, les alignements et les manques d’expériences au sein du domaine expérimental. Ils sont particulièrement adaptés au contexte numérique pour un objectif de production d’un modèle de substitution ou pour de l’optimisation, sur des modèles numériques complexes et non-linéaires. On peut regrouper dans cette catégorie les plans suivants :

- les plans basés sur des suites à faible discrédance, telles que les suites de Halton [Halton 60], de Sobol’ [Sobol’ 67], de Faure [Faure 82], de Hammersley [Hammersley 60] ou de Niederreiter [Niederreiter 88]. La discrédance rend compte de l’irrégularité de la répartition des expériences dans l’espace de conception [Franco 08]. Ces plans peuvent générer des motifs et alignements d’expériences lorsque le nombre de facteurs devient trop important. Cela peut être corrigé par un algorithme de scrambling [Vandewoestyne 06] ;
- les plans basés sur des critères de distance, telles que les critères maximin et minimax [Johnson Mark 90]. Les plans WSP [Beal 14], quant à eux, sont issus d’un processus de transformation d’un plan initial par élimination des expériences ;
- Le plan de Strauss est un plan basé sur un critère de répulsion suivant un paramètre de proximité. Tout comme les plans WSP, un plan initial est transformé, mais les expériences sont ici écartées les unes des autres ;
- l’hypercube latin, couramment utilisé [McKay 79], partage des propriétés avec les plans orthogonaux et peut être amélioré en appliquant des critères de type maximin

et minimax ;

- les plans à entropie maximale sont axés sur la maximisation de la quantité d'information apportée par chacune des expériences [Franco 08].

FIGURE 1.8 – Exemple d'un PEN de type Halton à 10 expériences utilisé pour le modèle T1

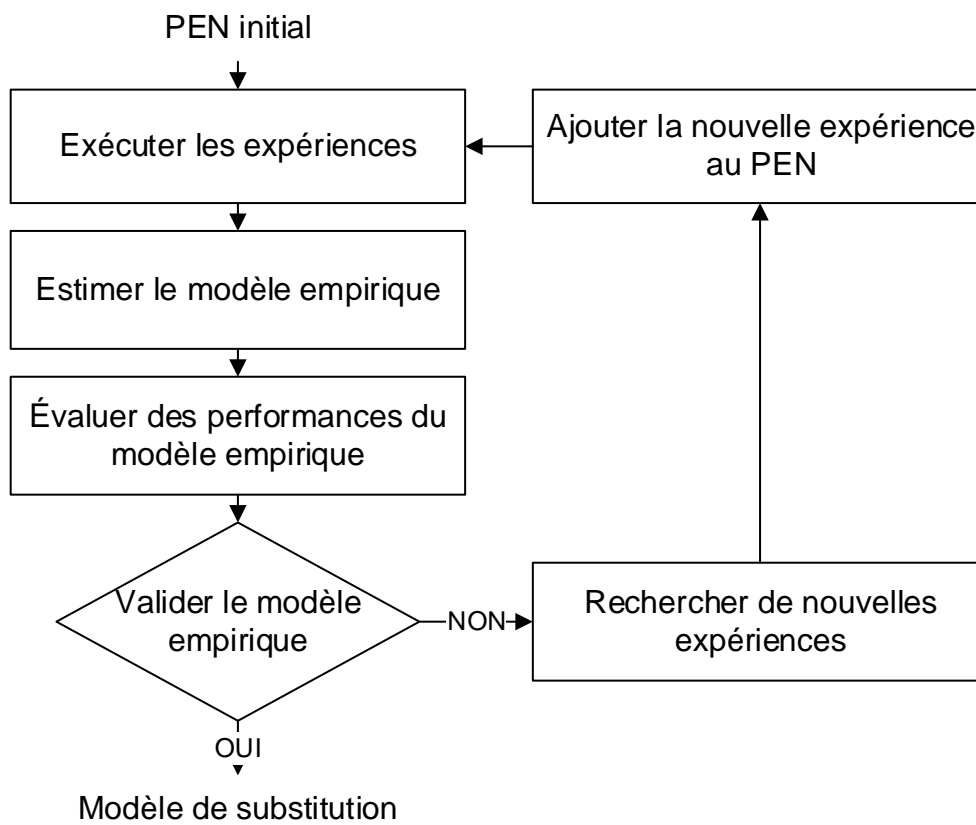


### Plans d'expériences adaptatifs

La mise en place d'un plan d'expériences impose la définition *a priori* des expériences. Une fois un premier PEN réalisé, il peut être nécessaire de l'enrichir avec d'autres expériences si certaines informations se révèlent être manquantes après analyse des résultats. Par exemple, si un modèle empirique résultant d'un PEN n'est pas assez précis, plutôt que de définir et exécuter un nouveau PEN, il est possible de rajouter quelques expériences pour améliorer le modèle empirique, sans ré-exécuter toutes les expériences précédentes. Certains types de plans décrits précédemment ont la capacité d'être complétés sans perdre leurs autres propriétés. Il est aussi possible de définir de nouvelles expériences pendant l'exécution du processus de PEN, selon des critères se focalisant sur l'amélioration des résultats de l'analyse du PEN plutôt que sur des propriétés du plan d'expériences. Les PEN basés sur la stratégie adaptative ont été utilisés pour les objectifs de production de modèle de substitution et d'optimisation [Iuliano 13, Dubourg 11, Kleijnen 12, Hu 08]. La Figure 1.9 présente l'algorithme général pour la production de modèle de substitution. Dans ce cas, un premier modèle empirique est estimé à partir d'un PEN initial. Tant que le modèle empirique ne peut être utilisé comme modèle de substitution, les expériences améliorant le mieux les performances du modèle empirique sont estimées, ajoutées au PEN et exécutées. A chaque itération, un nouveau modèle empirique est estimé, jusqu'à atteindre les niveaux de précision et de prédictivité désirés.



FIGURE 1.9 – Exemple de processus de PEN adaptatif pour la production d'un modèle de substitution



La mise en place de cette démarche nécessite le choix d'un algorithme d'optimisation (sous-section 1.2.7) et d'un critère d'optimalité. L'utilisation d'une stratégie adaptative n'est rentable que lorsque le coût et/ou le temps d'exécution d'une expérience sont très élevés et qu'aucun type de plans existant n'est adapté au problème traité. Cette stratégie permet d'obtenir le nombre d'expériences juste nécessaire pour atteindre l'objectif fixé. Le plan initial perd ses propriétés (uniformité, orthogonalité, etc.) au fur et à mesure que les expériences sont ajoutées. Le plan final est alors spécifique à l'application pour laquelle il a été calculé, et ne peut pas, *a priori*, être ré-utilisé dans un autre contexte.

Les critères d'optimalité, ou critère de remplissage du plan, dépendent de l'objectif poursuivi. Pour l'obtention d'un modèle de substitution, on cherche à minimiser l'erreur entre le modèle empirique et le modèle numérique initial. Une stratégie consiste à identifier le point du domaine expérimental présentant la plus grande erreur et de l'ajouter au PEN initial. Le critère d'Expected Improvement [Kleijnen 12] est d'avantage utilisé pour l'objectif d'optimisation s'appuyant sur un modèle de substitution.

### Synthèse sur les types de plans d'expériences

La définition d'un PEN nécessite de nombreux choix, et notamment le choix du type de plan d'expériences. Ce type conditionne la répartition des expériences dans le domaine expérimental. Il existe de nombreux types, dont certains nécessitent la définition de paramètres internes (ex : plans de Strauss, WSP) [Franco 08]. Le choix du type de plan dépend de l'objectif de l'étude, des ressources disponibles, du modèle empirique associé, etc. Mais de manière générale, ces choix ne répondent pas à une liste exhaustive de règles logiques. Il existe des travaux donnant des recommandations de choix [Chen 06, Sanchez 12, Schimmerling 17], mais de nombreux éléments doivent être définis selon le contexte spécifique de l'étude. Une fois le plan défini et les expériences réalisées, les méthodes d'analyse sont définies et appliquées. Le choix de ces méthodes dépend en partie de la définition du PEN. Ces méthodes sont détaillées dans les sous-sections suivantes.

#### 1.2.5 Méthodes d'estimation des modèles de substitution

La production d'un modèle de substitution vise à créer un nouveau modèle du modèle numérique initial, plus économique et plus rapide à évaluer. Bien que, historiquement, le modèle polynômial soit le plus répandu dans le domaine des plans d'expériences, de nombreux autres types de modèles existent [Wang 07]. Cette sous-section présente en premier lieu les notions d'apprentissage statistique liées à la production d'un modèle de substitution et le lien avec les PEN. Différents types de modèles empiriques, particulièrement étudiés lors de ces recherches, sont ensuite présentés : les modèles polynômiaux, les

réseaux de neurones et le Krigeage. Ces modèles sont étudiés pour identifier les différentes problématiques de choix, dans l'objectif de réaliser les choix les plus adaptés à un contexte donné.

## Apprentissage statistique

L'apprentissage statistique<sup>10</sup> est une discipline du domaine de l'Intelligence Artificielle. Dans le contexte de ces travaux, seules les méthodes d'apprentissage supervisé [Philip Chen 14] sont considérées. Ce domaine regroupe un ensemble de méthodes d'analyse des données pour en prédire de nouvelles.

Les **PEN** fournissent un ensemble de données d'apprentissage, chacune formée d'une expérience et de la réponse correspondante du système. Les propriétés de cet ensemble (quantité de données, diversité, exactitude, etc.) conditionnent directement la qualité du modèle empirique final. L'apprentissage du modèle empirique sur ces données d'apprentissage est assuré majoritairement par des algorithmes de régression ou d'intégration. Le modèle empirique peut être utilisé comme un modèle de substitution si [Dreyfus 11, Saporta 11] :

1. le modèle empirique est précis. La précision décrit l'adéquation aux données d'apprentissage. Lorsque le modèle est évalué sur une expérience d'apprentissage, l'erreur par rapport à la réponse du modèle numérique doit être la plus faible possible. Toutefois, cette erreur doit rester supérieure à une éventuelle erreur de mesure expérimentale. Dans le cas contraire, le modèle empirique serait plus précis que les données d'apprentissage, ce qui peut-être un symptôme du phénomène de sur-apprentissage (ou sur-ajustement) ;
2. le modèle empirique est prédictif. La prédictivité décrit la capacité de généralisation. Lorsque le modèle empirique est évalué sur une nouvelle expérience, l'erreur par rapport à la réponse du modèle numérique doit être la plus faible possible.

Le problème principal de l'apprentissage réside dans le "dilemme biais-variance". Trop augmenter la précision du modèle conduit à une baisse de sa prédictivité. Le modèle empirique est alors exposé au phénomène de sur-apprentissage : il reproduit très exactement les résultats des données d'apprentissage, mais commet une erreur très importante dès qu'il est évalué sur de nouvelles données.

Le contrôle de la précision et de la prédictivité est réalisé à l'aide de techniques de validation croisée (Figure 1.10) [Dreyfus 11], comme par exemple, la méthode du **Leave-One-Out (LOO)**, utilisée dans ces travaux. L'idée générale de la validation croisée est de partitionner l'ensemble des données à disposition en, au moins, deux ensembles : l'ensemble

---

10. ou apprentissage automatique, Machine learning en anglais

d'apprentissage et l'ensemble de validation <sup>11</sup>. Le modèle empirique est donc entraîné sur l'ensemble d'apprentissage. Une fois l'apprentissage terminé, la précision est évaluée en comparant les réponses observées contenues dans l'ensemble d'apprentissage et les réponses du modèle empirique sur les mêmes données d'apprentissage. La prédictivité est, quant à elle, évaluée par l'utilisation de l'ensemble de validation. On compare les résultats observés avec les résultats estimés par le modèle empirique, pour chaque donnée de l'ensemble de validation. Cette étape permet de quantifier la capacité de généralisation du modèle empirique sur des nouvelles données.

Si le modèle empirique est suffisamment précis et prédictif, il peut être utilisé comme modèle de substitution. Sinon, il est nécessaire de vérifier, et de modifier si besoin, les données d'apprentissage et de validation (quantité, exactitude, données aberrantes, diversité), et le type de modèle.

La méthode **LOO** produit  $n$  modèles empiriques entraînés chacun sur un ensemble d'apprentissage différent comportant  $n-1$  données. En notant  $y_i$  le résultat de l'expérience  $i$ ,  $\bar{y}$  la moyenne des résultats sur les  $n$  expériences,  $y_p$  le résultat prédit par le modèle empirique sur l'expérience de validation, on définit l'erreur  $\epsilon_i = y_i - y_p$  et les deux indicateurs suivants :

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n \epsilon_i^2}}{\bar{y}} \quad (1.1)$$

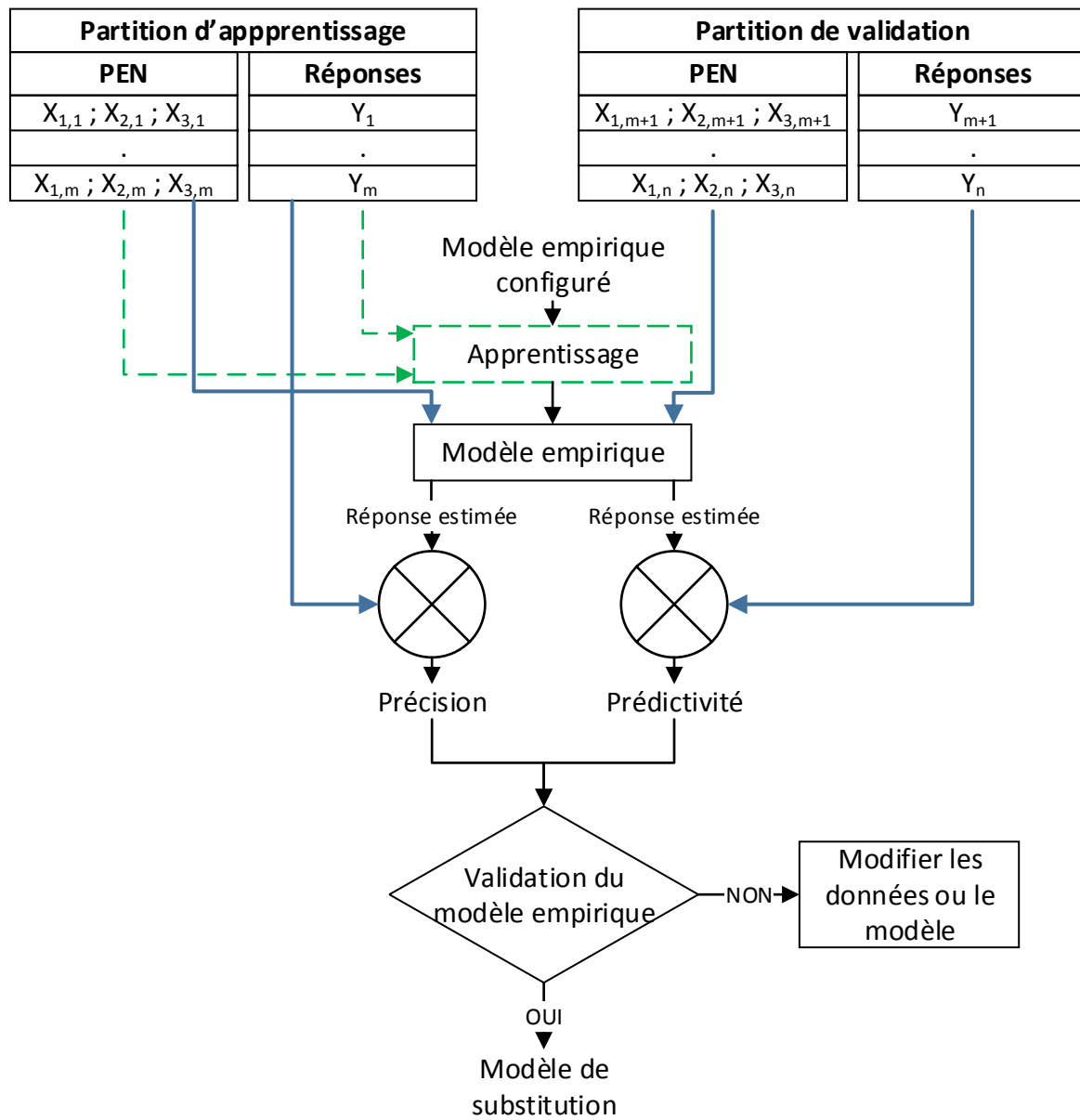
$$Q^2 = 1 - \frac{\sum_{i=1}^n \epsilon_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.2)$$

L'indicateur NRMSE (Normalised Root Mean Square error) [Cui 16a] permet de quantifier la précision du modèle empirique (équation 1.1). Plus il est proche de 0, plus la précision est élevée. L'indicateur  $Q^2$  [Iooss 14] permet de quantifier le niveau de prédictivité du modèle empirique (équation 1.2). Plus il est proche de 1, plus le modèle est prédictif. Il peut être interprété comme décrivant la part de variance expliquée par le modèle empirique. L'objectif est alors de minimiser NRMSE et maximiser  $Q^2$ . D'autres indicateurs existent, notamment lorsque les modèles à valider sont fortement non-linéaires [Iooss 14]. Ces deux indicateurs sont utilisés par la suite pour les modèles de chaos polynômial et de krigeage.

---

11. On peut trouver dans la pratique des partitionnements plus complexes, comprenant par exemple un ensemble de test inconnu de l'utilisateur qui mène l'opération d'apprentissage [Hastie 09]

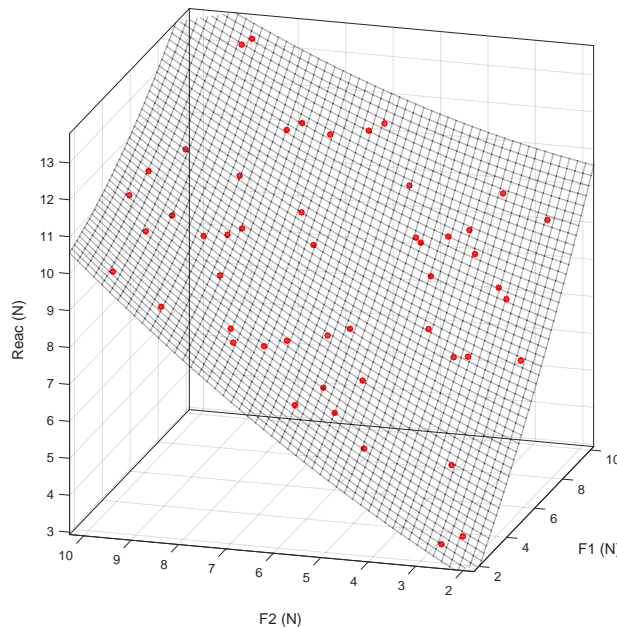
FIGURE 1.10 – Processus d'apprentissage pour la production d'un modèle de substitution avec validation croisée



## Types de modèles empiriques

**Les modèles polynomiaux** Les modèles polynomiaux sont les plus couramment utilisés en association avec les plans d'expériences. Une fonction polynomiale de degré  $p$  à  $q$  facteurs nécessite la réalisation de  $n = \frac{(p+q)!}{p!q!}$  expériences pour déterminer les coefficients  $\alpha_i, i = 1 \dots m$ . Ces coefficients sont estimés par la méthode des moindres carrés. Les polynômes permettent de visualiser rapidement les influences des facteurs et des interactions, mais sont réservés à la modélisation de comportements faiblement non-linéaires (Figure 1.11).

FIGURE 1.11 – Surface de réponses du modèle T1 (Reac=f(Fx;Fy)) obtenue par un plan de Halton et un polynôme de degré 2



Plus évolué, le modèle de chaos polynômial [Wiener 38, Cameron 47, Xiu 02] est une combinaison linéaire de polynômes orthogonaux, en considérant les facteurs comme des variables aléatoires. Ils sont capables d'approcher des fonctions plus complexes que les polynômes classiques. Un modèle de chaos polynômial d'ordre  $p$  est tel que :

$$y = f(\theta) = \sum_{k=1}^p \beta_k \cdot \psi_k(X(\theta))$$

avec :

- $\psi_k = \prod_{i=1}^q \varphi_{k,i}$ , où  $q$  est le nombre de facteurs,  $\varphi_{k,i}$  est un polynôme d'ordre  $k$  associé au facteur  $i$ , dont le type est choisi selon la loi de probabilité suivie par le facteur  $i$ . Si le facteur  $i$  suit une loi normale,  $\varphi_{k,i}$  sera le polynôme d'Hermite d'ordre  $k$  associé au facteur  $i$  [Crestaux 09];
- $\beta_k$ , le coefficient associé à  $\psi_k$ ;

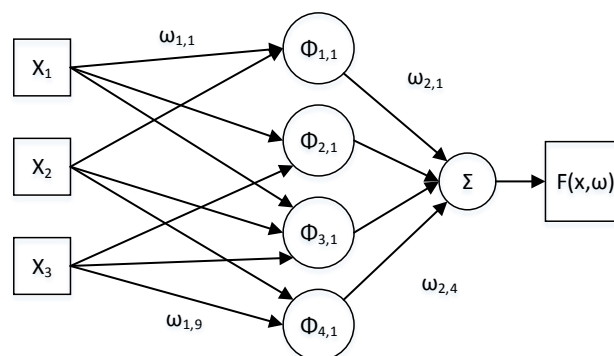
—  $X(\theta)$ , le vecteur des niveaux des facteurs (variables aléatoires dépendantes de  $\theta$ ).

Le nombre d'expériences nécessaires pour l'estimation des coefficients répond à la même règle que pour les polynômes classiques. Ces coefficients sont estimables par différentes méthodes, comme la méthode des moindres carrés ou par intégration numérique [Crestaux 09]. Les modèles de chaos polynômial présentent un avantage particulier : ils intègrent par nature l'estimation des indices de sensibilité de Sobol' (détaillé dans la sous-section 1.2.6). Les indices de Sobol' se déduisent directement à partir des coefficients calculés, et permettent de quantifier l'influence de chaque facteur et interaction sans calcul supplémentaire. En règle générale, les modèles polynômiaux ne peuvent pas approcher les comportements fortement non-linéaires.

**Les réseaux de neurones** Les réseaux de neurones sont utilisés notamment pour réaliser des surfaces de réponses complexes. Un réseau de neurones est une structure de plusieurs couches de neurones artificiels reliées entre elles. Un neurone est défini par des entrées, une fonction d'activation, et des sorties. Les fonctions d'activation peuvent être de différents types (fonction seuil, sigmoïde, etc.) [Dreyfus 11]. Une connexion entre deux neurones représente un coefficient de pondération  $\omega_i$ . Les coefficients  $\omega_i$  sont estimés par la méthode des moindres carrés, en utilisant un algorithme de rétropropagation. Dans le cadre de ces travaux, seuls les réseaux de type Perceptron Multi-Couches (PMC) [Shahin 14, Dreyfus 11] sont abordés. Ce type de réseau est constitué (Figure 1.12) :

- d'entrées, dont chaque élément représente un facteur du plan ;
- de  $n$  couches cachées composées de  $m$  neurones ;
- d'une dernière couche comprenant les neurones de sorties.

FIGURE 1.12 – Exemple de Perceptron Multi-Couches (1 couche cachée à 4 neurones, 3 entrées)



La Figure 1.12 représente un PMC à 1 couche cachée à 4 neurones et 3 entrées. En posant  $\phi$  la fonction d'activation,  $X$  le vecteur des niveaux des facteurs et  $\omega$  les coefficients

du modèles, la réponse  $F$  de ce PMC s'écrit :  $F(X, \omega) = \omega_2 \cdot \phi(\omega_1 \cdot X)$

Les PMC permettent d'approcher des fonctions fortement non-linéaires, mais l'apprentissage peut être coûteux. Un inconvénient est l'impossibilité d'estimer directement, contrairement au chaos polynômial, l'influence des facteurs et des interactions sur la réponse finale.

**Le krigeage** Les modèles de krigeage ont été développés initialement pour le domaine de la géo-statistique, afin d'estimer, dans toute une zone, les concentrations de minerais dans le sol à partir de quelques prélèvements en des endroits spécifiques [Matheron 67]. Ces modèles modélisent la réponse étudiée par une variable aléatoire, suivant une loi normale caractérisée par une moyenne  $\mu$  et une matrice de covariance donnée. La fonction de corrélation modélise la variabilité de la réponse  $Y(x)$  entre les expériences et ne dépend que de la distance entre le point d'estimation et les expériences entourant ce point. La réponse estimée est un estimateur linéaire transitif, sans biais, à variance minimale. L'estimation est donc exacte sur les expériences du plan d'apprentissage. L'erreur de prédiction est minimisée et quantifiée directement pour toutes les estimations. La transitivité permet l'ajout d'une expérience d'apprentissage sans modifier les prédictions ailleurs que dans le voisinage direct de la nouvelle expérience. L'objectif est de déterminer les coefficients  $\lambda_j$  pour minimiser la variance de cet estimateur. Il existe plusieurs types de krigeage [Baillargeon 05, Sluiter 09]. Dans le cas du krigeage universel, l'estimateur s'écrit :

$$\tilde{Y}(x) = \sum_{j=1}^n \lambda_j \cdot f_j(x)$$

, avec  $n$  expériences initiales, et  $f_j(x)$ , une fonction de tendance ne dépendant que de la distance entre le point d'estimation et les expériences du plan.

Il est nécessaire de choisir le type de la fonction de tendance (linéaire, quadratique, etc.), le type de fonction de corrélation (Gauss, exponentielle, etc.), la méthode de détermination des paramètres internes des fonctions de corrélation (méthodes de quantification de la variance et algorithme d'optimisation).

### Synthèse sur les types de modèles empiriques

L'obtention d'un modèle de substitution permet de remplacer le modèle numérique initial par un nouveau modèle rapide et peu coûteux. Les performances prédictives d'un modèle empirique dépendent du type de modèle, de la définition de ses paramètres internes, des données d'apprentissage provenant d'un [PEN](#) et du comportement étudié. De nombreux



types de modèles empiriques existent, et certains disposent de plusieurs paramètres internes à définir *a priori*. Ces choix, tout comme le choix des PEN, ne répondent pas entièrement à des règles bien définies et requièrent dans certains cas une expertise avancée ou de nombreux essais. Les modèles de produits complexes, non-linéaires et comprenant de nombreux facteurs, rendent ces choix plus difficiles, dans le but d'obtenir le meilleur modèle de substitution avec le moins d'expériences possible. La section suivante traite des méthodes permettant de réduire la complexité d'un problème en réduisant le nombre de facteurs.

### 1.2.6 Méthodes d'analyse de sensibilité

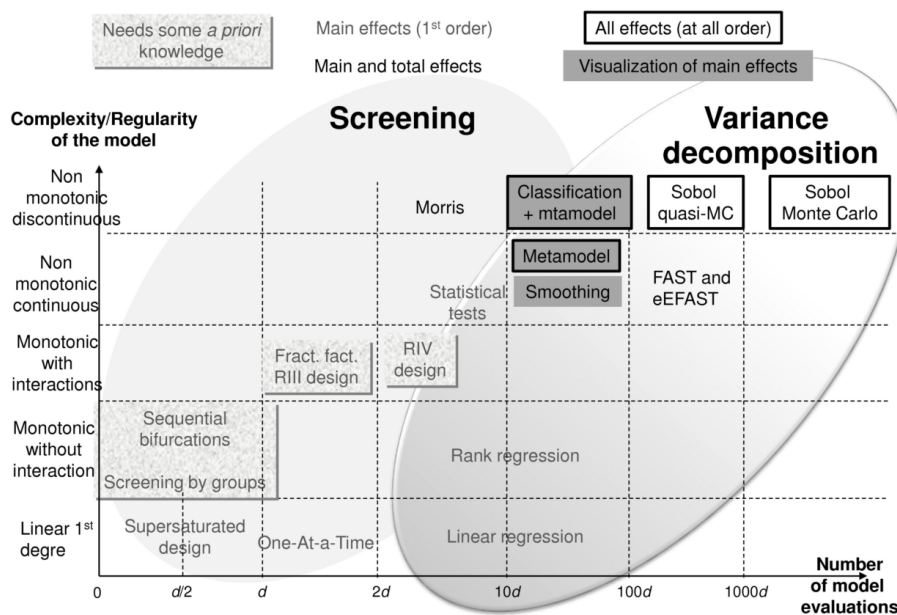
Il existe de nombreuses méthodes permettant de décrire qualitativement ou quantitativement la sensibilité d'une réponse par rapport aux facteurs du PEN. Ces méthodes sont classées en trois catégories [Saltelli 02, Iooss 14, Sergent 09] :

1. les méthodes de criblage pour une analyse qualitative. Ces méthodes ne requièrent que peu d'expériences et permettent de filtrer les facteurs et interactions n'ayant aucune influence. On retrouve dans cette catégorie la méthode de Morris, basé sur un plan "OAT" (One-At-Time) pour estimer les effets élémentaires des facteurs. On peut également citer l'utilisation de plans basés sur les matrices d'Hadamard ou de matrices supersaturées, les techniques de criblage par groupes et de bifurcation séquentielle [Sergent 09] ;
2. les méthodes locales sont basées sur le gradient de la réponse par rapport aux facteurs, autour d'une valeur nominale. L'estimation des sensibilités est réalisée par la méthode des différences finis ou par la méthode des états adjoints ;
3. les méthodes globales [Chastaing 13, Jacques 05] s'intéressent à l'effet de la variance des facteurs sur la variance de la réponse. Il existe plusieurs indices de sensibilité, quantifiant une mesure de la part de variance de la réponse expliqué par facteur donné (ou une interaction). On peut citer les indices SRC (Standard Regression Coefficient) et PCC (Partial Correlation Coefficient) pour les modèles linéaires, SRRC (Standardized Rank Regression Coefficient) et PRCC (Partial Rank Correlation Coefficient) pour les modèles non-linéaires monotones et, enfin, les indices de Sobol' [Iooss 14]. L'évaluation des indices de Sobol' peut s'avérer coûteuse. Il est alors courant de se baser sur un modèle de substitution, pour ensuite l'évaluer un grand nombre de fois et estimer ces indices. La qualité de ces indices est alors liée à la qualité du modèle de substitution. Le modèle de chaos polynômial peut être privilégié pour les indices de Sobol' : ils se déduisent directement des coefficients du modèle. La Figure 1.14 présente les indices de Sobol' d'ordre 1 et d'ordre 2 pour le modèle T1. Enfin, les méthodes d'analyse de la variance (i.e. analyse de la variance,

ou ANOVA<sup>12)</sup> s'appuient sur des tests d'hypothèses statistiques pour quantifier l'influence des facteurs.

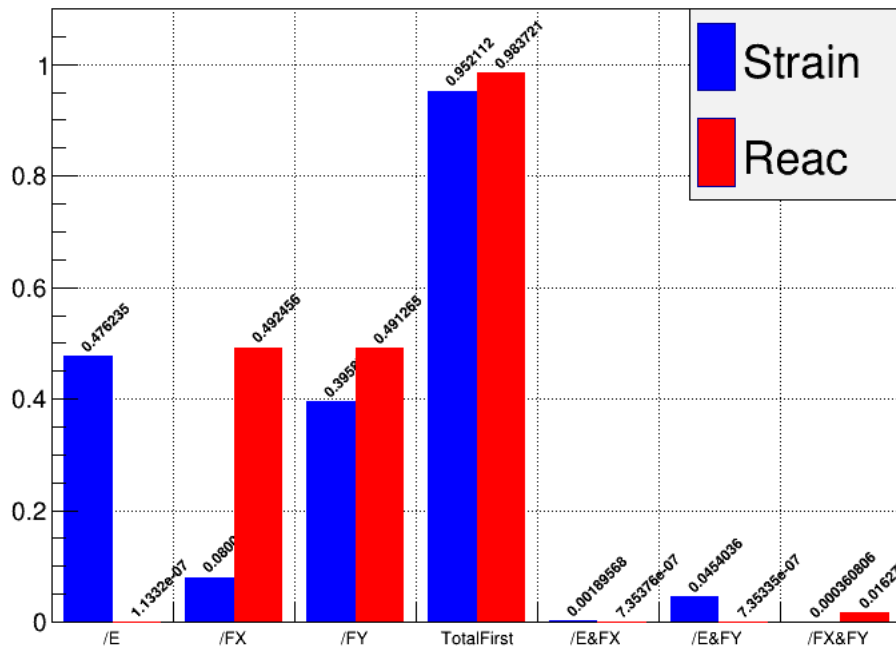
Le choix d'une méthode d'analyse de sensibilité dépend principalement de la nature du comportement du modèle numérique (non-linéarité, monotonie), du nombre de facteurs et du coût de chaque expérience. [Iooss 14, de Rocquigny 08] présentent un diagramme de choix simple permettant de choisir la méthode la plus adaptée à un problème donné (Figure 1.13). La définition du PEN, et du type de modèle empirique s'il est utilisé pour l'estimation des influences, impactent directement la qualité de l'estimation des indices, et donc les décisions prises pour négliger certains facteurs et en conserver d'autres. Une stratégie efficace est de mener plusieurs processus de PEN successifs. Un premier processus réalisant un criblage avec peu d'expériences, jusqu'au dernier comportant moins de facteurs et permettant l'estimation d'un modèle de substitution et d'indices de sensibilité quantitatifs précis.

FIGURE 1.13 – Synthèse des méthodes d'analyse de sensibilité selon le type de comportement du modèle numérique. Tirée de [Iooss 14].



12. pour ANalysis Of Variance

FIGURE 1.14 – Indices de Sobol’ pour le modèle T1 obtenus avec un modèle de substitution de chaos polynômial, basé sur un PEN de type Halton à 100 expériences, avec Uranie



Exemple d'interprétation : la variabilité de la réponse Strain est causée à 47.6% par le facteur E. Plus de 95% de la variabilité de la réponse Strain est due à des influences du premier ordre, ce qui implique que les interactions entre facteurs sont responsables de moins de 5% de la variabilité de Strain. Ces interactions peuvent être négligées.

### 1.2.7 Algorithmes d'optimisation

Les algorithmes d'optimisation sont utilisés à plusieurs niveaux au cours d'un processus de **PEN**. Ils peuvent être mis en œuvre pour optimiser le produit, mais aussi pour réaliser l'apprentissage d'un modèle empirique ou pour ajouter une expérience pour un **PEN** adaptatif. L'optimisation peut s'opérer à partir du modèle numérique initial ou d'un modèle de substitution préalablement produit.

Il existe de très nombreux algorithmes d'optimisation. Leur objectif principal est la recherche de l'optimum global d'une ou plusieurs fonctions objectifs sous certaines contraintes. Un algorithme d'optimisation se caractérise par sa capacité à prendre en compte les contraintes, sa capacité à être utilisé pour des problèmes multi-objectifs, son coût d'exécution, sa vitesse de convergence vers l'optimum et sa tendance à converger vers des optima locaux ou vers l'optimum global. Pour mener une démarche d'optimisation, il est nécessaire de trouver un compromis entre l'exploration du domaine admissible et l'intensification des calculs pour estimer précisément l'optimum. Se concentrer sur l'intensification donnera un optimum précis, mais sans assurance que cet optimum soit bien l'optimum global.

La première catégorie comprend des méthodes locales. On peut citer la méthode du simplexe, l'algorithme de Newton, de quasi-Newton ou encore la méthode du Lagrangien augmenté [Allaire 05]. Certains de ces algorithmes reposent sur le calcul du gradient de la fonction, voir même de la matrice hessienne, ce qui occasionne un coût de calcul élevé. Un **PEN** peut être utilisé pour fournir un ensemble de points de départ pour exécuter une optimisation plusieurs fois en parallèle et ainsi améliorer l'exploration du domaine.

La seconde catégorie comprend des méthodes globales, les métaheuristiques [Boussaïd 13]. Ces méthodes reposent sur des algorithmes inspirés de fonctionnement de systèmes naturels. Cette catégorie comprend, entre autres, les algorithmes évolutionnaires (ex : algorithme génétique), et des méthodes d'intelligence en essaim (ex : algorithmes d'essaims particuliers, de colonies de fourmis). Les métaheuristiques ont une meilleure tendance à converger vers l'optimum global. Les algorithmes évolutionnaires basés sur les essais disposent d'une population explorant le domaine admissible et intensifiant les calculs en plusieurs zones du domaine simultanément. Ces algorithmes sont donc parallélisables et peuvent proposer un ensemble de solutions candidates. Ils sont également adaptés aux problèmes multi-objectifs. Toutefois, la vitesse de convergence de ces algorithmes est difficilement maîtrisable. Le choix de leurs paramètres internes est donc une tâche difficile (choix de la taille de la population par exemple), et souvent réalisée selon une expertise métier.

### 1.2.8 Synthèse : Le processus de plan d'expériences numériques

Le processus de **PEN** est utilisé pour différents objectifs menant à un raccourcissement de la phase de conception, une meilleure maîtrise du produit et à une qualité accrue. Dans le cas où le processus de **PEN** est basé sur un modèle numérique coûteux, il devient nécessaire de minimiser le nombre d'expériences requises pour atteindre un objectif particulier. Cette minimisation repose sur l'utilisation d'un ensemble de méthodes de planification d'expériences et d'analyse des résultats, complexes et diversifiées. De nombreux choix sont nécessaires pour configurer entièrement et efficacement le processus de **PEN**, selon le coût et la qualité des résultats et des analyses.

Bien que des guides et les recommandations de choix existent [Schimmerling 17], ils restent généraux et ne peuvent être exhaustifs. Ces travaux nous ont menés à participer à la rédaction de la nouvelle version du fascicule de documentation FD-X06-081, en collaboration avec de nombreux acteurs industriels (Renault, PSA, Valeo, Nexter, SNECMA) et universitaires (UTC, Universités Aix-Marseille). Ce guide, fruit d'une collaboration soutenue d'acteurs de l'industrie et de chercheurs, apporte une aide à la mise en œuvre des plans d'expériences et également des **PEN**. De nombreuses règles régissent ces choix, mais le contexte d'utilisation reste un élément essentiel à la définition complète des plans d'expériences.

Certaines méthodes ne répondent pas entièrement à des règles claires et explicites, et le problème de choix dépend toujours d'un contexte d'utilisation particulier. De plus, chaque choix peut impacter directement le choix d'autres éléments de la configuration du processus de **PEN**. Par exemple, le choix de la méthode d'analyse de sensibilité peut dépendre du type de méthode de production d'un modèle de substitution, qui dépend lui-même de la définition du **PEN**.

Les choix de configurations deviennent de plus en plus complexes et cruciaux suivant la complexité, le coût et le temps d'exécution du modèle numérique initial. Une simulation multi-physique, comportant plusieurs dizaines de facteurs, simulant un comportement très complexe et occasionnant un temps d'exécution élevé, peut nécessiter plusieurs processus de **PEN** successifs et l'utilisation de méthodes complexes. Par exemple, une ou plusieurs analyses de sensibilité peuvent filtrer les facteurs et interactions non-influents, suivies d'un processus de **PEN** adaptatif pour la production d'un modèle de substitution basé sur un réseau de neurones performant. Une nouvelle analyse de sensibilité basée sur les indices de Sobol' peut être menée pour quantifier précisément les influences des facteurs et interactions restants. Enfin, une démarche d'optimisation peut être mise en place pour optimiser le produit.

Chacune de ces démarches tend à réduire le coût et le temps d'exécution nécessaires pour

la prise en compte des variabilités. Pourtant, plus ces démarches sont précises et efficaces pour traiter des problèmes complexes, plus leur mise en application requiert une expertise avancée. De plus, de nombreuses possibilités sont offertes aux utilisateurs. Ces derniers ne disposent pas nécessairement de l'expertise requise pour déterminer la configuration de chaque processus de **PEN** leurs permettant d'obtenir les résultats souhaités à un coût et à une durée d'exécution minimaux. Les utilisateurs peuvent alors être amenés à effectuer plusieurs essais.

Ce manque d'expertise conduit à une perte de temps, soit par une phase de recherche et des essais infructueux, soit par le refus d'utiliser des outils trop complexes qui pourraient conduire à un gain de temps futur. Un processus de **PEN** offre de nombreux avantages qui peuvent se révéler inaccessibles pour des produits complexes. Une capitalisation et une réutilisation de l'expertise de l'entreprise pourrait accélérer la phase de configuration du processus de **PEN** pour appliquer ces méthodes plus rapidement et plus efficacement.

## 1.3 Gestion des données pour la simulation numérique

Les sections précédentes ont présenté l'importance des processus de simulations numériques et de **PEN**, ainsi que de l'application d'une stratégie **PLM** pour assurer la compétitivité des entreprises de l'industrie mécanique. Il a été relevé, entre autre, une complexité élevée de ces processus et une tendance à produire de très grandes quantités de données. Cette section aborde la problématique de gestion des processus de simulation et de gestion des données.

### 1.3.1 Gestion du cycle de vie des simulations

Les enjeux de l'application d'une stratégie spécifique aux activités liées à la simulation numérique ont été relevés par [CIMDATA 11] :

- ces processus entraînent une forte croissance des données à gérer. Ce phénomène est dû à la croissance des capacités de traitements et de gestion des données (réseaux rapides et puissance de calcul croissante), au besoin de simuler des comportements complexes, et à une forte hausse de la confiance attribuée dans les prédictions des modèles numériques. La croissance de ces volumes de données reste supérieure à la montée en puissance des moyens informatiques [CIMDATA 12] ;
- l'utilisation de ces processus peut entraîner des pertes de temps rallongeant la phase de conception sans apporter de valeurs. Le manque d'intégration des experts en simulation dans le reste du processus de développement, le manque de partage de données et de connaissances, le manque d'automatisation et de réutilisation des modèles numériques ou des processus déjà mis en place rallongent le processus de développement des produits par des opérations manuelles répétitives et des re-développements, de modèles numériques par exemple, coûteux ;
- la capitalisation des bonnes pratiques en entreprise est identifiée comme un facteur clé pour faire évoluer ses produits, innover et maintenir sa compétitivité face à ses concurrents.

Ces trois enjeux majeurs ont conduit à considérer une stratégie globale de gestion des processus de simulations numériques.

Le projet européen CRESCENDO [CRESCENDO 13], rassemblant des entreprises d'aéronautique et spatial (Airbus, Aircelle, SNECMA, Thales, Rolls Royce) et de développement de logiciels de simulation (Ansys, Samtech, Dassault Systèmes, etc.) et de nombreuses universités, a proposé de très nombreuses solutions pour accroître la qualité des

simulations, réduire leurs coûts et leurs temps d'exécution dans un milieu collaboratif. Le but de ce projet a été de proposer une maquette numérique intégrant le comportement et la définition de tous les composants d'un avion ("Behavioural Digital Aircraft") disponible pour toutes les parties prenantes de l'entreprise étendue. Parmi les innovations proposées, on peut citer un environnement collaboratif unifié, le support des processus de simulations multi-physiques, la traçabilité des données et des processus, la possibilité de ré-utiliser les données, les connaissances, les processus, ou encore l'intégration des processus de vérification et de validation des simulations au plus tôt.

La gestion du cycle de vie de la simulation ([Simulation Lifecycle Management \(SLM\)](#)) est issue du besoin de conserver la propriété intellectuelle acquise par la pratique de simulations numériques, de gérer et de diffuser les données et connaissances, d'automatiser les processus et d'ajouter le comportement du produit dans la maquette numérique du produit. La stratégie [SLM](#) est bâtie sur quatre objectifs :

1. la gestion des données de simulation et d'essais expérimentaux. Cette fonctionnalité lie les données de simulations et les données expérimentales aux autres données du produit, dont les exigences du client, les exigences réglementaires et la maquette numérique. Cela permet, en plus d'améliorer la traçabilité, le partage et la réutilisation des données, d'accélérer et de capitaliser les prises de décisions de validation du produit. Enfin, les différentes versions des modèles de simulation, liées à des essais successifs ou des versions et configurations de produits différentes, sont également prises en compte ;
2. la gestion des processus de simulation et d'essais expérimentaux. Le [SLM](#) gère les exécutions des processus de simulation, pour toutes les disciplines et outils spécialisés et par rapport aux moyens de calculs à disposition (calculs locaux, clusters [HPC](#), etc.). L'application de cette fonctionnalité est à adapter en fonction des capacités de l'entreprise et de la nouveauté du type de simulation considérées. Le processus de simulation peut être réalisé manuellement ou de façon totalement automatisée, sans interaction avec l'utilisateur. Cela permet également de contrôler les activités de simulation et de capitaliser les processus pour constituer des modèles de processus réutilisables ;
3. l'aide à la décision. Les résultats des processus de simulation sont présentés de façon à rendre les décisions prises meilleures et plus rapides. Les processus de [PEN](#) et d'optimisation, impliquant de nombreuses décisions quant à l'exploration des solutions possibles et à la qualité du produit, sont également concernées ;
4. le support des processus collaboratifs et la sécurisation des données, à l'image de la stratégie [PLM](#).



### 1.3.2 Gestion des données de simulation

La gestion des données de simulation (**SDM**) est définie comme une technologie basée sur des bases de données pour gérer les données des simulations et des processus de simulation tout au long du cycle de vie du produit [Norris 12]. Les données à gérer sont multiples : la définition du modèle géométrique utilisé (différent du modèle **CAO**), la définition des matériaux, des phénomènes, des cas de simulation, la définition des divers paramètres du modèles, les exigences liées à la simulation (performance attendue du produit étudié), la définition du modèle en lui-même (maillage, algorithmes utilisés, etc.), les résultats de simulations et des essais expérimentaux.

Les difficultés rencontrées par les bureaux d'étude [Norris 12] concernent la traçabilité des données, leur accessibilité parmi toutes les données disponibles et la représentation pertinente de ces données pour accélérer les prises de décisions. Le contexte collaboratif apporte trois difficultés supplémentaires. Premièrement, les différents bureaux d'études impliqués peuvent être éloignés géographiquement, rendant le transfert des données de simulation (de l'ordre du Go ou du To pour les modèles les plus complexes) lent et difficile. Deuxièmement, les modèles multi-physiques ou multi-échelles peuvent poser des problèmes de gestion des versions des modèles pour l'exploitation des résultats, et d'interopérabilité dû à la multiplicité des outils utilisés dans la chaîne de calcul. Les résultats peuvent alors être sous plusieurs formats différents et dépendre d'un ensemble de modèles chaînés, dont il faut contrôler la version. Troisièmement, il est dans certains cas nécessaire de considérer le besoin pour chaque entreprise collaboratrice de conserver sa propriété intellectuelle. Cela impose une limitation des transferts des données selon l'origine et le rôle de chaque personne impliquée.

En termes d'applications, le rapport publié par la NAFEMS (National Agency for Finite Element Methods and Standards) [Norris 12] présente en particulier l'exemple du constructeur automobile Audi. L'entreprise a appliqué un système **SDM** pour gérer près de 1000 simulations par jour (analyse de crash, acoustique, vibration, etc.). Le système rassemblait, au moment de la publication de ces chiffres, près de 500000 processus de simulation, chacun générant en moyenne 200 documents, répartis dans 14 disciplines et concernant 450 collaborateurs. Sa productivité a été augmentée de 35%. Ces chiffres illustrent les enjeux industriels pour la gestion des données et des processus spécifiquement adaptés aux processus de simulations numériques.

### 1.3.3 Synthèse sur la gestion des données de simulation

L'application d'une stratégie **SLM** présente de grands avantages pour les entreprises ayant fait de la simulation un élément clé pour réduire les coûts et le temps de développement

tout en améliorant la qualité des produits. Une utilisation intensive de la simulation numérique produit de grands volumes de données hétérogènes, d'autant plus grands qu'ils concernent des produits complexes nécessitant la collaboration de plusieurs entités. Par la stratégie [SLM](#), les données spécifiques des activités de simulation sont capitalisées, au même titre que les données de conception par exemple. Cela contribue à la mise en place d'une maquette numérique du produit intégrant le comportement du produit, d'accélérer les processus de décision et d'en garantir la traçabilité. Les processus de [PEN](#), basés sur une utilisation intensive des outils de simulation, sont directement concernés par l'application du [SLM](#). Cette stratégie peut supporter la mise en place et l'exécution des processus de [PEN](#) en capitalisant les données de simulation, mais aussi, les données spécifiques aux [PEN](#).

## 1.4 Conclusion et problématique

Ce premier chapitre a permis de relever trois points principaux liés au développement de produit.

La première section a présenté les grands enjeux de l'industrie mécanique, et notamment l'intérêt que représente l'application de la simulation numérique dans le processus de conception, ainsi que l'application d'une stratégie PLM. Ces deux solutions apportent des outils pour maintenir la compétitivité des entreprises de l'industrie mécanique, en mettant sur le marché des produits innovants, plus fiables, plus robustes, rapidement et à un coût minimal.

La seconde section s'est focalisée sur le processus de PEN, pour la prise en compte des variabilités dans les simulations numériques. L'application de ce type de processus pose des difficultés pour des cas industriels. D'une part, ce processus peut considérablement allonger le processus de développement du produit. Des méthodes existent pour palier à ce phénomène (production d'un modèle de substitution, analyse de sensibilité, planification adaptative). D'autre part, ces méthodes sont d'autant plus nécessaires que la complexité du produit s'accroît (calculs longs, nombreux paramètres, comportements non-linéaires). Le premier verrou repose sur le fait que ces méthodes sont également d'autant plus difficiles à appliquer efficacement que le produit est complexe. Le large éventail de choix et de paramétrages de ces méthodes nécessite une expertise avancée pour obtenir rapidement le processus de PEN le plus économe et le plus riche en interprétation pour un problème donné.

La troisième section a détaillé les solutions de gestion des données et des processus spécifiquement pour la simulation numérique. La stratégie SLM peut apporter un gain de productivité significatif aux entreprises. Le processus de PEN peut générer de très grands volumes de données hétérogènes qu'il est nécessaire de capitaliser. Le support des processus de PEN est implicite au sein de la stratégie SLM, dans le sens où ils ne sont considérés que comme des séries de simulations, et non pas comme un ensemble de méthodes complexes. Le second verrou porte sur la capitalisation des données propres aux PEN, en vue de leurs réutilisation.

La problématique considérée dans ces travaux se situe au croisement des enjeux de l'ingénierie collaborative, de la gestion du cycle de vie du produit, de l'utilisation de plus en plus fréquente des outils de simulations numériques et de la prise en compte des variabilités au sein même des modèles numériques. Le processus de PEN apporte des solutions pour améliorer la conception du produit et réduire le coût de sa propre exécution par la réduction du nombre d'expériences nécessaires. Toutefois, l'application des méthodes réduisant ce coût d'exécution exige une expertise avancée et, pour certaines méthodes (e.g.

les PEN adaptatifs) des campagnes d'essais, pour configurer le processus de PEN lui-même. Cette difficulté tend à annuler l'économie offerte par l'utilisation de ces méthodes lors de l'exécution du processus de PEN. L'autre aspect du processus de PEN est qu'il génère de grandes quantités de données. La problématique de ces travaux porte sur le moyen de capitaliser et de réutiliser ces données et l'expertise existante pour accélérer la mise en place du processus de PEN. Cela permettra de rendre le processus de PEN applicable plus facilement à des cas industriels complexes en assistant les concepteurs dans leurs choix, de façon à réduire la durée du processus de développement de produits en économisant à la fois des ressources de calcul et du temps de configuration du processus de PEN.

L'état de l'art, présenté dans le chapitre suivant, s'intéresse aux problématiques de capitalisation, de partage et de réutilisation des données et des connaissances pour accélérer simultanément la phase de configuration et la phase d'exécution du processus de PEN. Cet état de l'art est orienté d'une part sur la structuration des connaissances pour les capitaliser au travers d'une entreprise étendue et, d'autre part, sur la réutilisation des connaissances capitalisées pour fournir une aide à la configuration des processus de PEN.

# Chapitre 2

## État de l'art

■ *"Mundi placet et spiritus minima", ça n'a aucun sens mais on pourrait très bien imaginer une traduction du type : "Le roseau plie, mais ne cède... qu'en cas de pépin" ce qui ne veut rien dire non plus.* ■

---

François Rollin (Le roi Loth), Kaamelott, Livre VI, Lacrimosa, écrit par A.Astier.

### Sommaire

---

<b>2.1</b>	<b>Gestion des connaissances</b>	<b>53</b>
<b>2.2</b>	<b>Modèles de données et de connaissances</b>	<b>58</b>
2.2.1	Modèles de données	58
2.2.2	Modèles de connaissances	65
2.2.3	Synthèse sur les modèles de connaissances	72
<b>2.3</b>	<b>Réutilisation des connaissances</b>	<b>74</b>
2.3.1	Critères d'analyses	76
2.3.2	Méthodes de réutilisation	77
2.3.3	Comparaison des méthodes de réutilisation	85
<b>2.4</b>	<b>Synthèse</b>	<b>86</b>

---



Le chapitre précédent a présenté les enjeux actuels de l'industrie mécanique, et plus particulièrement les axes de développements et verrous concernant l'utilisation de la simulation numérique durant le développement du produit. Le besoin de prendre en compte les variabilités au sein des modèles de simulation a été identifié, de même que le besoin de gérer les données et les connaissances tout au long du cycle de vie du produit.

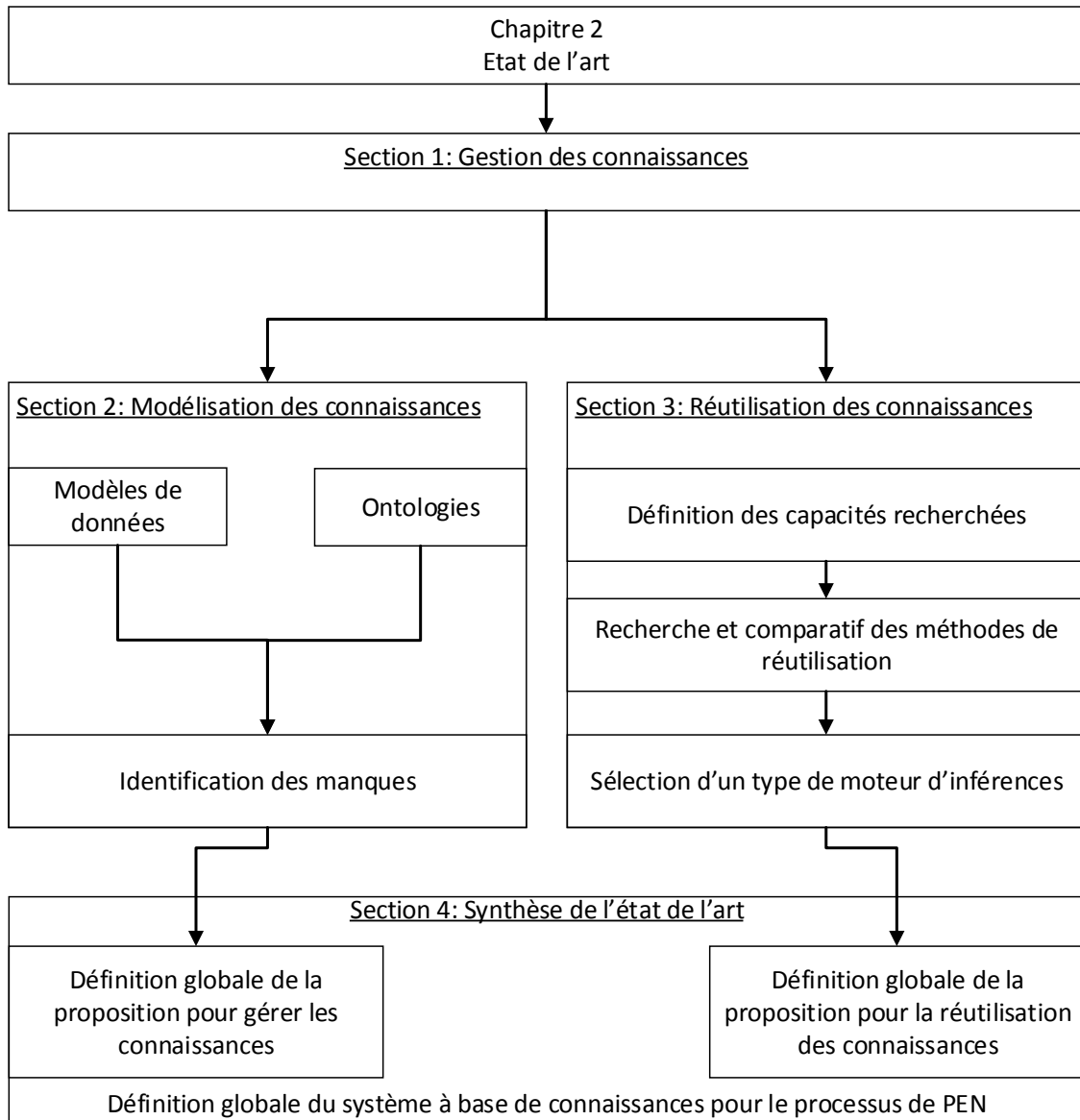
Basé sur le processus de simulation numérique, le processus de **PEN** accroît de façon significative les besoins en termes de gestion de données. De plus, l'application de processus de **PEN** à des produits complexes nécessite l'utilisation de méthodes exigeant des connaissances expertes avancées. Sans une expertise appropriée, l'application de ces méthodes peut se révéler contre-productive. En effet, le temps nécessaire à la configuration peut dépasser le temps potentiellement gagné par l'application de ces méthodes. Pis, une mauvaise configuration de ces méthodes peut augmenter le temps et le coût d'exécution du processus de **PEN**, tout en diminuant l'interprétabilité des résultats. Dans un contexte d'entreprise étendue, l'ensemble des acteurs concernés ne dispose pas nécessairement de ces connaissances, d'autant plus que certaines des méthodes employées sont fondées sur un comportement aléatoire et dépendant du produit développé.

L'intégration du processus de **PEN** dans une démarche **PLM** pourrait permettre d'améliorer la capitalisation, la traçabilité, le partage et la réutilisation des données et des connaissances propres à ce processus. Cette intégration répond à des besoins de garantir la traçabilité des décisions, processus et produits, de réduire le temps et le coût de développement, d'assurer l'amélioration continue du produit et de pouvoir développer des produits complexes dans un contexte d'ingénierie collaborative et d'entreprises étendues.

Ce chapitre dresse un état de l'art décomposé en quatre sections (Figure 2.1) :

- la section 1 définit les concepts principaux de la gestion des connaissances ;
- la section 2 se focalise sur la modélisation et structuration des connaissances pour favoriser leur capitalisation et l'échange. Cette partie de l'état de l'art couvre les modèles de données et les modèles ontologiques liés à la conception de produit de façon générale et aux plans d'expériences en particulier ;
- la section 3 se concentre sur les moyens de réutilisation des connaissances pour assister le concepteur dans ses choix. Cette section compare plusieurs méthodes de l'**Intelligence Artificielle (IA)** selon des critères définis au préalable. La méthode d'analyse des connaissances sera alors déterminée ;
- la section 4 synthétise l'état de l'art pour présenter la proposition visant à améliorer la gestion et la réutilisation des connaissances relatives au processus de **PEN**, afin de réduire les temps de configuration et d'exécution.

FIGURE 2.1 – Plan du chapitre 2





## 2.1 Gestion des connaissances

La connaissance constitue une ressource importante pour la réussite d'une entreprise. La gestion de ces connaissances au sein de l'entreprise réponds à plusieurs enjeux, identifiés par [Dalkir 05] :

- le contexte d'entreprise étendue induit des problématiques de travail collaboratif [Robin 07], tels que l'échange des connaissances et la compréhension commune ;
- le besoin de compétitivité et d'innovation (objectifs de coût, de qualité et de délai) nécessitent une capitalisation et une réutilisation des connaissances de l'entreprise. Cela accélère le développement de produits en évitant de devoir refaire un travail déjà effectué, ou de redécouvrir des connaissances perdues. Une base de connaissances est alors disponible pour, par exemple, améliorer le produit ;
- la nécessité de ne pas perdre les connaissances expertes et de constituer une mémoire de l'entreprise. Le départ d'un collaborateur s'accompagne d'une perte de ses connaissances pour l'entreprise si cette capitalisation n'est pas réalisée ;
- la progression des technologies informatiques génère une accélération des flux d'informations. Cela impose une réactivité de plus en plus forte de la part de chaque collaborateur, mais apporte des moyens pour mieux partager les connaissances.

De nombreuses définitions de la gestion de connaissances existent [Kebede 10, Sarrafzadeh 06, Cheung 10]. L'état de l'art de ce chapitre se base sur une définition présentée par [Sarrafzadeh 06] : "The creation and subsequent management of an environment which encourages knowledge to be created, shared, learnt, enhanced, organized for the benefit of the organization and its customers". La gestion de connaissances couvre ainsi la capitalisation, la transmission et l'apprentissage, la réutilisation, l'amélioration et la création de connaissances au sein de l'entreprise.

La définition de la connaissance se base sur plusieurs distinctions.

La première distinction repose sur la différence entre la connaissance, l'information et la donnée [Chandrasegaran 13]. La donnée peut prendre la forme d'observations, de quantités factuelles. La notion de donnée peut être définie comme l'expression d'un fait quantifiable par une valeur [Owen 02]. L'information est une donnée interprétée, contextualisée. La connaissance, quant à elle, résulte de l'expérience acquise par une personne à partir de ces informations.

La seconde distinction est à faire entre les connaissances explicites et les connaissances tacites [Nonaka 01]. Les connaissances explicites sont facilement exprimables et transmissibles par des informations et des données, par le biais de documents écrits, par exemple. Les connaissances tacites sont liées à l'expertise et au savoir-faire personnel, acquis durant

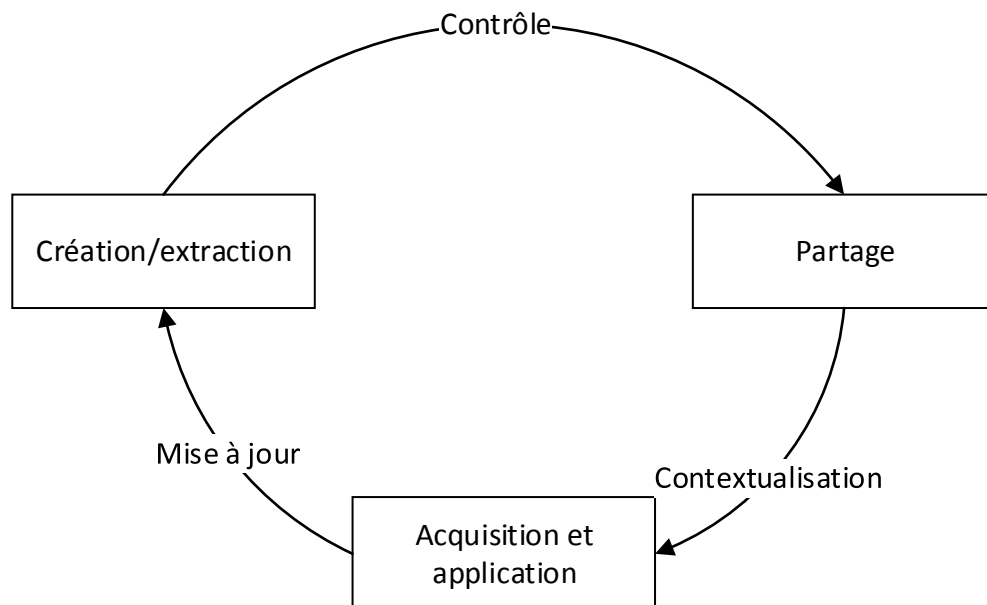
de longues périodes d'apprentissage et de pratique. La connaissance tacite est difficile à transmettre et à capitaliser.

D'autres distinctions existent, telles que la distinction entre les connaissances compilées (issus de l'expérience, donnant des résultats pour un contexte donné) et les connaissances dynamiques (théories, équations, etc., permettant de générer d'autres connaissances), ou la distinction entre les connaissances liées au produit, et celles liées aux processus [Chandrasegaran 13].

La gestion des connaissances est modélisable par un cycle. [Dalkir 05] propose un modèle de cycle de gestion de connaissances synthétisant les précédents développements effectués. Ce modèle comprend trois étapes principales (Figure 2.2) :

1. la création et/ou l'extraction des connaissances ;
2. les connaissances sont contrôlées et peuvent ensuite être diffusées ;
3. les connaissances sont contextualisées pour être acquises et utilisées par d'autres collaborateurs. Elles sont alors mises à jour après utilisation, ce qui crée de nouvelles connaissances.

FIGURE 2.2 – Cycle synthétique de la gestion de la connaissances, tiré de [Dalkir 05]



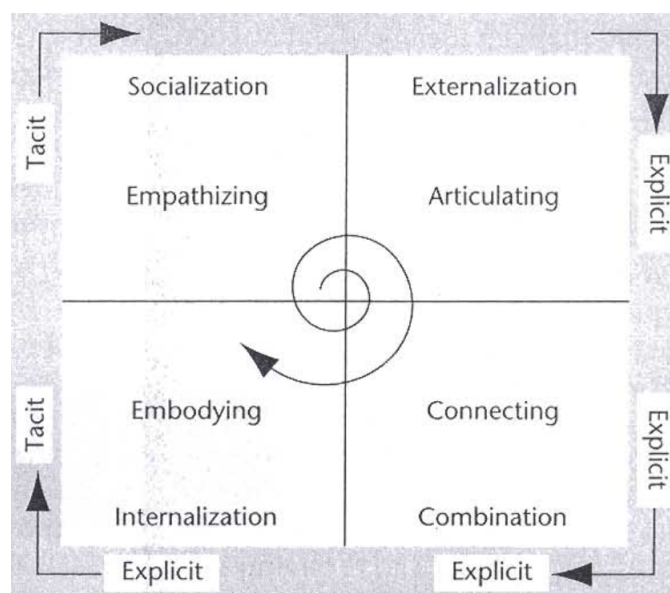
Les connaissances personnelles de chaque collaborateur sont alors amenées à être utilisées, communiquées et capitalisées. La communication et l'apprentissage de connaissances nécessitent une conversion entre les connaissances tacites et explicites. [Nonaka 01] ont modélisés ces conversions en quatre étapes, représentées dans la Figure 2.3 :

1. la socialisation consiste à transmettre directement d'une personne à une autre des connaissances tacites, par la discussion ou l'observation par exemple ;

2. l'externalisation est la communication de connaissances tacites en connaissances explicites, qui conduit à créer de nouvelles connaissances capitalisables par l'entreprise, par le biais de rapports écrits ou de supports de présentation ;
3. l'étape de combinaison est une transformation de connaissances explicites pour les représenter d'une nouvelle manière ;
4. l'internalisation est le passage de connaissances explicites en connaissances tacites, qui peut se manifester par l'acquisition de savoir-faire par l'expérience, la lecture de documents, etc.

Ces quatre étapes forme un cycle de transformation. Ce cycle concerne tous les acteurs de l'entreprise : les collaborateurs, les équipes, les services, les départements, etc.

FIGURE 2.3 – Modèle de conversion des connaissances, tirée de [Nonaka 01]



L'état de l'art présenté dans les sections suivantes repose sur ce cycle de transformation continue des connaissances et s'est focalisée sur deux aspects de la gestion des connaissances pour l'aide à la configuration des processus de [PEN](#) :

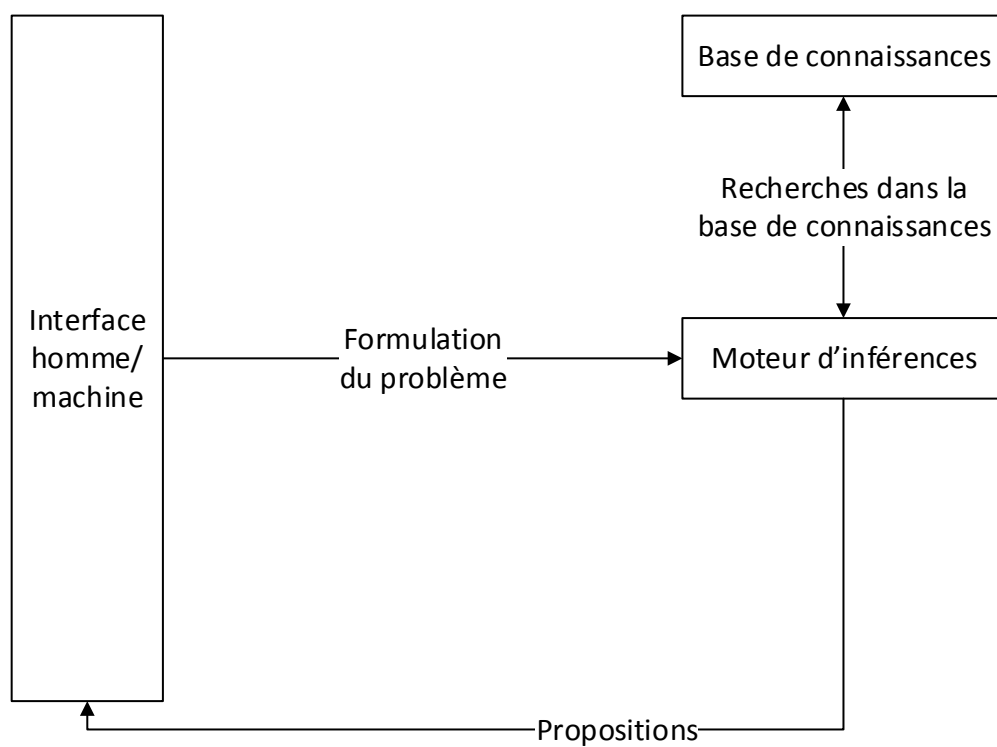
1. la structuration des connaissances pour apporter un moyen de capitalisation, de partage et de diffusion au sein de l'entreprise. Une des solutions pour supporter la capitalisation et le partage des connaissances consiste à modéliser, structurer et organiser ces connaissances par le biais d'un modèle ontologique. Une ontologie peut apporter une modélisation commune adaptée au contexte collaboratif. En particulier, les nombreuses possibilités de configuration d'un processus de [PEN](#) peuvent être représentées de manière compréhensible pour tous les acteurs concernés, non-experts du domaine ;
2. la réutilisation des connaissances capitalisées peut être assurée par certaines méthodes du domaine de l'[IA](#), comme par exemple, les systèmes experts et les techniques

d'apprentissages (Machine Learning). Dans ce dernier cas, la connaissance peut être créée à partir des données par une étape d'internalisation des connaissances explicites par les utilisateurs.

L'objectif de ces recherches est de mettre en place un système à base de connaissances pour apporter une aide à la configuration des processus de [PEN](#). Ce système vise à répondre à la demande de connaissances expertes pour la mise en place rapide de processus de [PEN](#) efficaces pour des produits complexes, en étant un support de capitalisation des connaissances des experts et une source de création et d'enrichissement des connaissances par l'analyse des données et par le partage des connaissances. L'architecture d'un tel système se décompose en trois parties principales [Shu-Hsien Liao 05] (Figure [2.4](#)) :

1. une base de connaissances, rassemblant les connaissances capitalisées et garantissant leur disponibilité. Cette base permet d'externaliser et d'internaliser des connaissances ;
2. un moteur d'inférences, raisonnant sur les données et connaissances capitalisées pour assister l'utilisateur. Le moteur d'inférences se place à la phase de combinaison de connaissances dans le modèle de conversion des connaissances (Figure [2.3](#)) ;
3. une interface homme/machine, permettant à l'utilisateur de soumettre au système son problème de configuration spécifique et de visualiser et comprendre les résultats du moteur d'inférences. Elle facilite la communication des connaissances entre le système et l'utilisateur, et supporte donc les différents types de conversion des connaissances, réalisées entre l'utilisateur, le moteur d'inférences et la base de connaissances.

FIGURE 2.4 – Structure globale d'un système d'aide à la décision à base de connaissances



## 2.2 Modèles de données et de connaissances

Cette section dresse un état de l'art sur la structuration des données et des connaissances liées à la conception de produit. La démarche de recherche s'appuie sur deux dichotomies :

- la recherche de modèles de données et de modèles de connaissances. La hiérarchie existante entre les données et les connaissances place les données comme étant une base pour l'expression et l'acquisition de connaissances. La recherche de modèles de connaissances peut se révéler infructueuse, mais cela ne signifie pas nécessairement qu'il n'existe pas de descriptions à l'échelle des données. Un modèle de connaissances peut s'appuyer sur une structuration de données existantes ;
- la recherche de modélisations générales et spécifiques. Le processus de **PEN** est dépendant des autres processus de développement (simulation, conception, etc.). Les données et les connaissances liées aux processus de **PEN** englobent des concepts généraux, par rapport au produit ou au projet, et des concepts spécifiques, concernant par exemple, le choix d'un type de plan et d'un type de modèle empirique en fonction du modèle de simulation et des ressources nécessaires.

### 2.2.1 Modèles de données

#### Modèles de données pour la conception et la fabrication

Une structuration des données liées aux processus de développement du produit permet de faciliter la gestion, la traçabilité et les échanges de données au sein d'une entreprise étendue. Cette partie couvre les principaux travaux de modélisation des données décrivant le produit et les processus associés.

[Fenves 06] propose un modèle de données générique pour la description d'un produit. Cette structuration des données vise à assurer la gestion de grandes quantités de données hétérogènes. Le "Core Product Model" (CPM) apporte des moyens de structurer, filtrer, intégrer et partager ces données de façon pertinente et efficace, pour supporter l'application d'une stratégie **PLM**. Le concept central de CPM est celui d'artefact ("artifact"), représentant les sous-ensembles du produit considéré. Chaque artefact est caractérisé par sa fonction (afin de respecter les exigences du cahier des charges), sa forme (définition physique pour remplir sa fonction) et son comportement (représentant le fonctionnement du produit pour remplir ses fonctions). CPM permet donc de capitaliser les données décrivant le produit en gardant les relations entre ces données. Chaque artefact est référencé en relation avec ses fonctions, sa définition physique, ses modes de fonctionnement, les exigences associées. CPM intègre également des données informatives, telles que des explications de choix de conceptions et les données relatives au processus de développement de produits.

Ce modèle très général peut être étendu [Fiorentini 07] et spécialisé à l'aide d'autres structures, telles que :

- "Open Assembly Model" [Fenves 03b], détaillant la définition géométrique, le comportement cinématique du produit et toutes les données de tolérancement ;
- "Master product Model" et "Design-Analysis Integration Model" [Fenves 03b, Fenves 03a], découlant d'une volonté de lier les modèles géométriques aux modèles de simulations idéalisés. Un modèle géométrique principal du produit est alors lié à toutes les idéalizations nécessaires aux activités de simulations et d'analyses ;
- le "Product Family Evolution Model" [Fenves 03b], permettant de prendre en compte les évolutions du produit. Cette extension assure la traçabilité entre les différentes versions et améliorations du produit. Les causes de ces évolutions sont capitalisées et associées à chaque version.

L'[International Organisation for Standardisation \(ISO\)](#) a spécifié la norme [ISO 10303 Standard for the Exchange of Product model data \(STEP\)](#) [Pratt 05], développée sous la responsabilité du sous-comité TC184/SC4 "Données industrielles"<sup>1</sup>, spécifie, entre autres, plusieurs protocoles d'application ([Application Protocol \(AP\)](#)). Ces [AP](#) définissent plusieurs modèles de données spécialisés dans un ou plusieurs domaines. La norme [STEP](#) est en constante évolution. Certains [AP](#) ont connu plusieurs éditions pour enrichir les modèles de données, améliorer l'interopérabilité entre les processus, améliorer la cohérence entre les [AP](#), etc.

L'[AP239](#) (édition 3, 2015) [ISO 15] intitulé "Product Life-Cycle Support" (PLCS) supporte l'échange des données du produit tout au long de son cycle de vie, et plus particulièrement durant les phases de conception et de production. Plus particulièrement, cet [AP](#) supporte notamment :

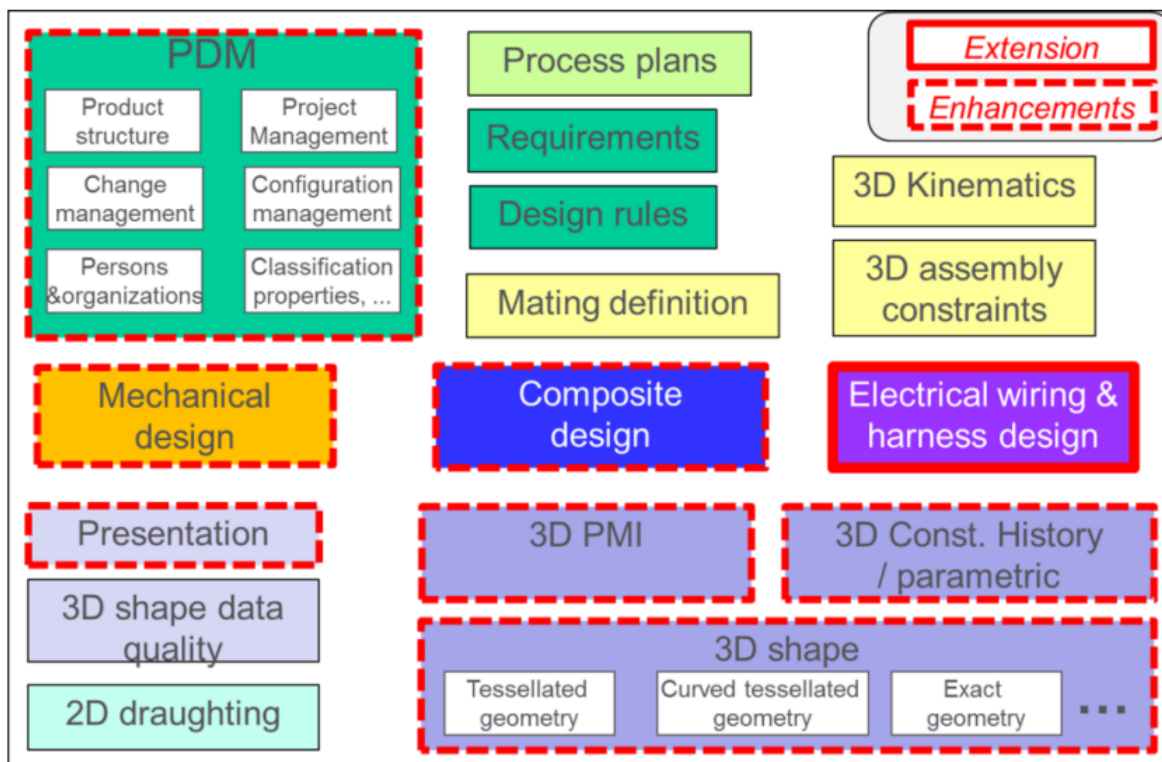
- les données de définition et de maintien de produits complexes ;
- les données permettant de gérer les changements, par les différentes configurations et versions du produits ;
- les données décrivant les assemblages, la structure du produit et les différents niveaux de décompositions. Ces données supportent l'identification et la représentation des pièces, composants, sous-systèmes et interfaces, mais aussi les versions, les fonctions et les documents associés à chaque composants et sous-systèmes ;
- la gestion des activités au sein des processus (spécification des tâches, des ressources, du planning, des compétences et des ressources humaines), permettant d'assurer la traçabilité des processus et des produits.

---

1. sous-comité 4 du comité technique 184 "Systèmes d'automatisation et intégration"

L'AP242 [ISO 14], "Managed model based 3D engineering" résulte du regroupement des AP203 "Configuration controlled 3D design" et AP214 "Core data for automotive mechanical design processes". Son objectif est d'intégrer toutes les données permettant de décrire le produit, en cohérence avec les AP 209 "Multi-Disciplinary Analysis and design", AP 210 "Electronic assembly, interconnect, and packaging design", AP 233 "Systems Engineering", AP 238 "Application interpreted model for computer numeric controllers" (STEP-NC) et AP 239 "Product Life-Cycle Support". Les différents types de données gérés par l'AP242 (édition 2), et les améliorations et ajouts par rapport à la première édition, sont représentés par la Figure 2.5.

FIGURE 2.5 – Représentation globale des fonctionnalités de l'AP242 édition 2, tirée de [Brangé 14]



L'ensemble de ces développements, débutés en 1984 pour améliorer l'échange des données géométriques en spécifiant un format neutre (le format STEP), dénote un besoin et une volonté de gérer et partager l'ensemble des données relatives au produit durant tout son cycle de vie.

L'existence de modèles de données normalisés offre une base pour la structuration des données et des connaissances liées aux produits et aux processus, et pour leurs compréhensions communes dans un contexte collaboratif. Ces modèles décrivent des données utilisées pour la définition et la traçabilité des processus de PEN, comme par exemple la définition du produit, de ses versions, de ses composants, mais aussi des exigences ou des workflows.



## Modèles de données pour simulations numériques

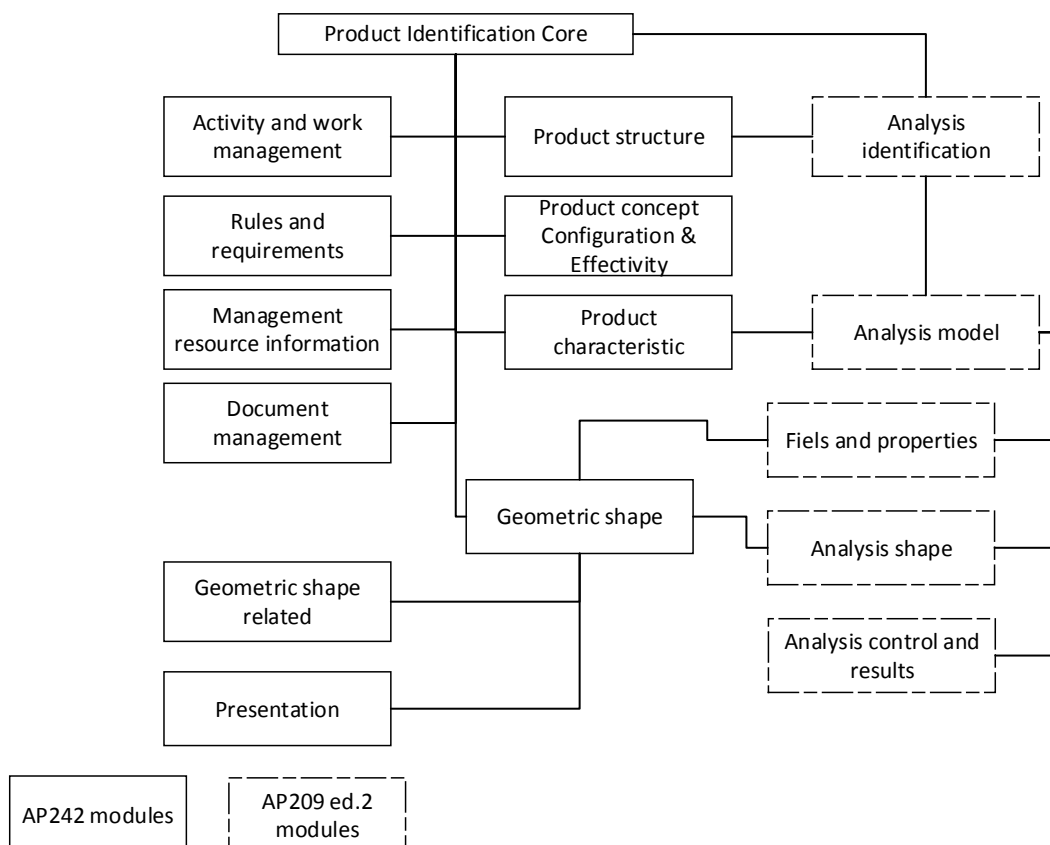
La croissance des activités de simulations numériques pour le développement de produits a conduit à l'intégration de la description des données spécifiques dans les modèles de données déjà définis pour le produit. Ces évolutions conduisent à la nécessité de gérer de grandes quantités de données produites par l'utilisation d'outils de simulation.

L'utilisation des processus de simulations par les entreprises vise à :

1. réduire le coût et le temps de développement, en prenant des décisions en amont de la phase de conception, et en réduisant le nombre de prototypes nécessaires à l'obtention d'un produit conforme aux exigences ;
2. accroître la qualité du produit par des modélisations précises des comportements, conduisant à une optimisation de sa conception.

L'AP 209, "Multi-Disciplinary Analysis and design" [ISO 10] spécifie une description très précise des données de simulation. L'AP 209 est construit en lien avec l'AP242, décrit précédemment, et ajoute une définition des données relatives à la définition des modèles numériques, notamment les modèles par éléments finis pour la mécanique des structures et des fluides (Figure 2.6).

FIGURE 2.6 – Architecture des données supportées par l'AP 209 édition 2 (adaptée de [ISO 10])



L'AP 209 ed.2 couvre notamment :

- la description du produit dans son aspect mécanique, en tant qu'assemblage de composants et de pièces. Les données de configurations, de versions et de changements au cours du cycle de vie sont également supportées ;
- les données géométriques et topologiques du produit, ainsi que les tolérances associées ;
- les données décrivant les matériaux utilisés, telles que leur composition chimique et leurs propriétés physiques. Une description détaillée des matériaux composites est aussi intégrée au modèle de données ;
- la définition des modèles géométriques détaillés (pour la CAO) et idéalisés pour le processus de simulation numérique ;
- la description des données du processus de simulations numériques, telles que les données relatives aux modèles par éléments finis, les types d'analyses en mécanique des solides et des fluides ;
- la gestion des hypothèses de modélisation et d'idéalisation ;
- les documents et les graphiques résultants.

Par exemple, le module d'application 1372 "Analysis" de l'AP 209 ed.2 détaille le concept de modèle numérique en fournissant les données suivantes :

- la description géométrique et comportemental du modèle numérique ;
- un identifiant, un nom, une description, un type d'analyse, les logiciels associés ;
- la description du maillage spatial ;
- l'identification de chaque exécution du modèle numérique ;
- les résultats de chaque exécution.

Concernant la simulation par la méthode des éléments finis, largement appliquée dans l'industrie mécanique, le modèle EXODUS 2 [Schoof 94] propose une description très précise du modèle de simulation. L'intégralité des éléments constituant le modèle numérique est détaillée : les types d'éléments, la définition des nœuds, les résultats, les propriétés matériaux et géométriques, etc. Ce modèle est limité en termes d'intégration au sein d'un processus de conception. Par exemple, il n'est pas rattaché à des données de traçabilité plus générales, ni aux ressources de calculs à disposition.

### Modèle de données pour les PEN

Les processus de PEN sont identifiés par [CIMDATA 12, CIMDATA 14] comme faisant partie des processus à gérer au sein d'un processus de simulations numériques. Toutefois, aucun modèle de données, modélisant ces processus, n'a été identifié au cours de la démarche de recherche engagée dans le cadre de ces travaux. Les données spécifiques, telles

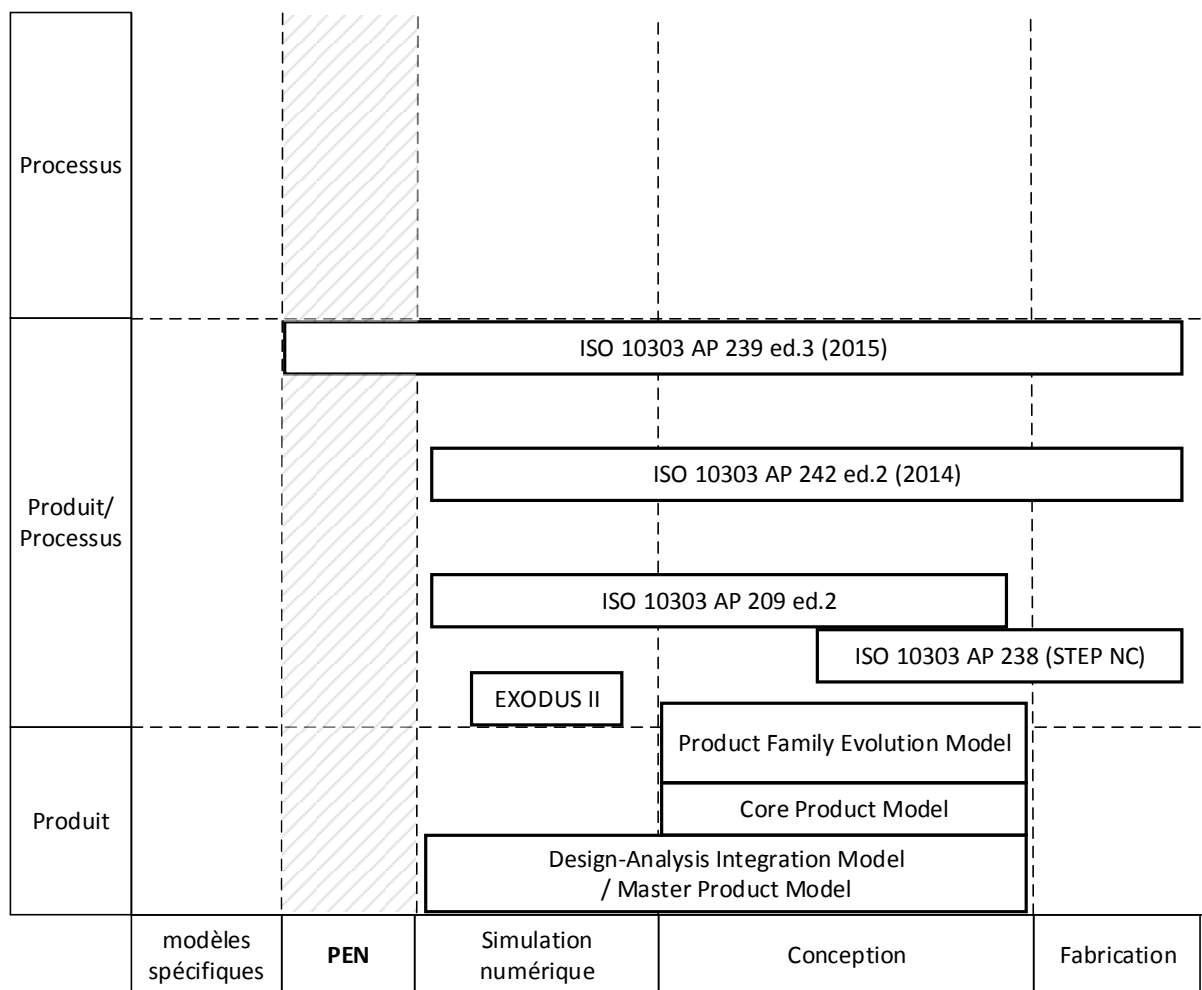
que les données techniques et paramétriques des processus de **PEN** ne profitent d'aucune structuration spécifique. La description des données de contexte est, en revanche, déjà supportée par les modèles de données détaillés dans les paragraphes précédents.

### **Synthèse des modèles de données**

Cette première recherche de structures de données a permis d'identifier essentiellement des structurations normalisées. Elles contribuent à la capitalisation, la gestion et au partage des données décrivant les produits et les processus de conception et de simulations. Pourtant, les données spécifiques aux processus de **PEN**, telles que les méthodes employées, ne sont pas décrites par ces modèles. Cette recherche a permis d'identifier un manque en terme de gestion de ces données spécifiques (Figure 2.7).

Le manque identifié doit être vérifié avec un éventuel manque en terme de description des connaissances. La partie suivante se concentre sur l'analyse des modèles de connaissances ontologiques. Si des ontologies sont identifiées comme gérant des données spécifiques aux processus de **PEN**, cela impliquera une gestion des données.

FIGURE 2.7 – Classification synthétique des modèles de données étudiées



## 2.2.2 Modèles de connaissances

Une manière de représenter les connaissances est de les modéliser sous la forme d'une ontologie. Cette approche est centrée sur le sens existant entre les différents concepts décrits. [Gruber 93a, Feilmayr 16] définissent une ontologie comme une spécification explicite d'une conceptualisation. Une ontologie vise à décrire un domaine de connaissance spécifique par une expression communément comprise et partageable.

La conception d'une ontologie repose sur quatre propriétés principales, énoncées par [Gruber 93a] :

- la clarté des définitions des concepts. La définition du domaine sémantique ne doit montrer aucune incohérence logique.
- la capacité d'enrichissement, par l'ajout de nouveaux concepts. Ces ajouts permettent une spécialisation de l'ontologie ou une extension de son domaine sémantique. Il est également possible de lier différentes ontologies pour assurer une continuité sémantique entre les différents développements.
- un biais d'encodage minimal. En d'autres termes, le sens des concepts et la compréhension par les utilisateurs ne doit pas dépendre, ou dépendre le moins possible, du mode d'expression de ces concepts ;
- une généricité maximale. La généricité renforce la compréhension commune.

Le respect de ces critères résulte d'un compromis. Une ontologie trop générique n'est pas applicable dans des domaines précis, mais une trop forte spécialisation limite la compréhension. Une ontologie trop spécialisée peut également conduire à des problèmes de clarté.

En pratique, une ontologie est composée de concepts centraux ayant une signification particulière, liés entre eux par des relations sémantiques. Un langage couramment utilisé pour expliciter un domaine sémantique est le langage [Ontology Web Language \(OWL\)](#) (Ontology Web Language), spécifié par le W3C (World Wide Web Consortium).

[OWL](#) est un langage permettant d'exprimer des connaissances. Ce langage est basé sur le modèle de données RDF (Resource Description Framework), et RDF-S (RDF-Schema) (Figure 2.8). RDF-S permet de créer des ontologies simples, peu expressives. [OWL](#), et son évolution OWL2, apporte une plus grande expressivité. De nombreux types de propriétés permettent de définir avec une meilleure exactitude le sens des concepts et des liens entre concepts. [OWL](#) permet de construire des axiomes et des propositions logiques, ajoutant une capacité de raisonnement.

La première étape de la construction d'une ontologie consiste à définir une structure sémantique. Cette structure est définie par des concepts principaux, représentés par des classes et par des propriétés liant les classes entre-elles. Les propriétés peuvent être des

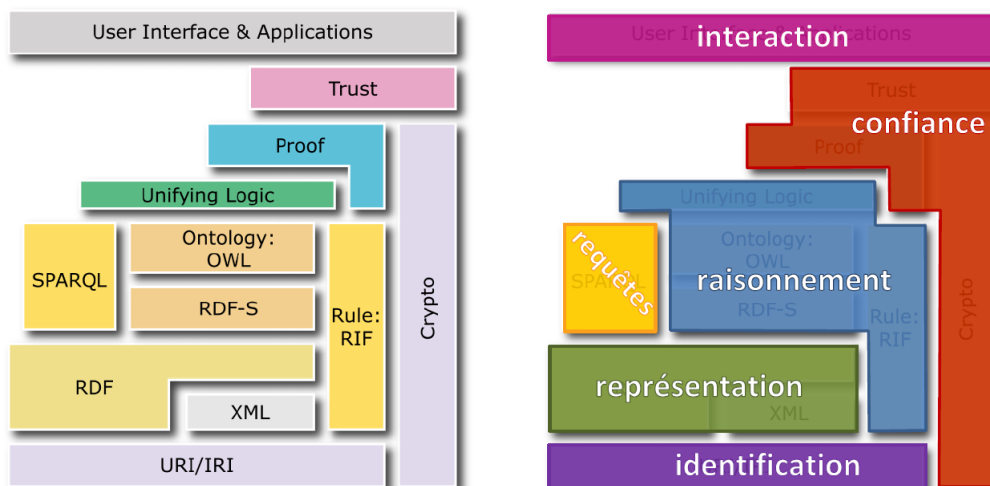
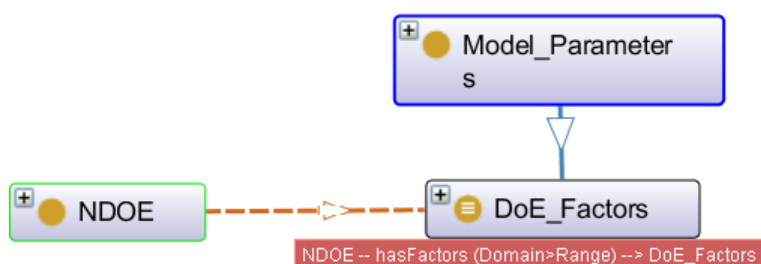


FIGURE 2.8 – Standards du web spécifiés par le W3C

propriétés "objet", des propriétés de données ou des annotations. Les propriétés de type "objet" permettent de lier les classes entre elles, alors que les propriétés de type "données" lient une classe à un type de données. La Figure 2.9 représente trois classes, liées par deux propriétés. La classe "DOE\_Factor" (représentant le concept de facteur de tous plans d'expériences) est une sous-classe de "Model\_Parameters". On décrit ainsi que les facteurs sont des paramètres d'un modèle numérique, mais que l'inverse n'est pas vrai. La classe "NDOE" (représentant le concept de PEN) est liée à la classe "DOE\_Factor" par la propriété "hasFactors". Ainsi, une instance de la classe "NDOE" sera liée à une instance de la classe "DOE\_Factor", permettant alors de retrouver les facteurs assignés à un PEN en particulier.

FIGURE 2.9 – Exemple de relations ontologiques



OWL permet d'enrichir le sens de cette propriété en ajoutant des caractéristiques. Le Tableau 2.1 présente certaines caractéristiques disponibles dans OWL et analyse leur sens dans le cas des classes présentées à la Figure 2.9. D'autres propriétés sont disponibles, permettant de définir, par exemple, des équivalences ou une cardinalité (ex : "un PEN a au moins un facteur").

Caractéristiques	Sens	Application
Fonctionnelle	Un PEN ne peut avoir qu'un seul et unique facteur	Non, il peut y avoir plusieurs facteurs par PEN
Fonctionnelle inverse	Si deux facteurs sont identiques, alors ils sont impliqués dans un seul et même PEN	Non, un facteur peut être utilisé dans plusieurs PEN différents
Transitive	Si un PEN a un facteur F1, et que F1 a un facteur F2, alors le PEN est lié au facteur 2	Non, cette propriété dépend d'autres classes. Pour certains types de plans, l'interdépendance entre facteurs n'est pas possible.
Symétrique	Si un PEN (P1) a un facteur F1, F1 a un facteur (P1)	Non, un facteur n'est pas un PEN
Asymétrique	Si un PEN (P1) a un facteur F1, le facteur F1 ne peut avoir un facteur (P1)	Oui

TABLEAU 2.1 – Exemples de caractéristiques d'une propriété objet : analyse pour le cas des concepts de PEN et de facteurs

La propriété de données permet de spécifier des valeurs associées aux instances de classes. Par exemple, dans la proposition "le **PEN** n°1 comprend 3 expériences", "le **PEN** n°1" est une instance de la classe "NDOE" et 3 est la valeur de la propriété de donnée "NumberOfExperiments", de type "entier positif".

Les annotations permettent de clarifier la définition des concepts en ajoutant une description. Ces descriptions sont caractérisées par la langue dans lesquelles elles sont exprimées. Cela permet d'améliorer la compréhension et la diffusion de l'ontologie.

**OWL** définit plusieurs profils en fonction de l'expressivité désirée et de la complexité des inférences. Les trois profils principaux sont, dans un ordre croissant d'expressivité, **OWL-Lite**, **OWL-DL** (Description Logic) et **OWL-Full**.

La seconde étape consiste à instancier l'ontologie, en y ajoutant des données selon la structure préalablement définie.

En troisième étape, l'ontologie est vérifiée par un moteur d'inférences. Les inférences réalisées à partir d'une ontologie permettent notamment de vérifier la cohérence de l'ontologie. La syntaxe est d'abord vérifiée, puis la cohérence logique dans la structure et entre la structure et les instances. L'expressivité maximale offerte par **OWL-Full** ne permet pas, dans tous les cas, de réaliser des inférences complètes avec les moteurs d'inférences actuels. Les moteurs d'inférences associés aux ontologies sont nombreux. Ces moteurs diffèrent suivant, notamment, les types de logiques qu'ils manipulent, leurs capacités à réaliser des inférences sur les classes et les instances, les types de propriétés supportés

et leurs performances [Abburu 12]. On peut citer, parmi les moteurs libres, HermiT<sup>2</sup>, FaCT++<sup>3</sup> ou encore Pellet<sup>4</sup>. Le moteur d'inférences associé à une ontologie permet également de déduire de nouvelles propriétés et d'expliquer le raisonnement. Ainsi, si une erreur est détectée, le moteur d'inférences peut guider l'utilisateur vers la source de l'erreur. De plus, si une nouvelle propriété est déduite, les raisons de cette déduction seront rapportées à l'utilisateur.

Une propriété fondamentale de **OWL** est sa définition reposant sur les standards du web. Une ontologie basée sur **OWL** est alors facilement accessible et partageable.

L'état de l'art des ontologies se rapportant aux processus de **PEN** est présenté dans les deux sections suivantes. La démarche de recherche de ces ontologies a été effectuée suivant deux approches convergentes. La première approche est la recherche d'ontologies décrivant les connaissances liées aux produits, et comprenant des concepts spécialisables pour être adaptées aux processus de **PEN**. La seconde approche consiste à rechercher les ontologies spécialisées dans les méthodes spécifiques aux processus de **PEN**. L'analyse de ces ontologies permet de définir une proposition d'ontologie pour lier les développements existants et les compléter.

## Ontologies produits

La première recherche s'intéresse aux ontologies supportant des concepts généraux, tels que les données relatives aux processus, aux produits. Les ontologies décrites dans cette partie ont été conçues dans le cadre d'applications dans le domaine de la conception mécanique, pour améliorer la capitalisation et l'échange des données et des connaissances entre les différents outils numériques et les différents processus. Ces ontologies sont analysées par rapport à leur capacité de support des données plus spécialisées, telles que les données relatives à la simulation numérique et à la méthode des **PEN**.

L'ontologie PRONOIA2 a été conçue pour gérer les données de produits tout au long du début du cycle de vie, comprenant les phases de conception et de fabrication [Gruhier 15]. Son originalité est d'aborder la gestion des changements et des transformations du produit par un concept de spatio-temporalité. Cette ontologie est composée de trois niveaux de spécialisation. Le premier niveau décrit les concepts les plus généraux et est considérée comme une méta-ontologie. Ce niveau comprend notamment les concepts de régions et de primitives (opérations), chacun détaillé en régions et primitives spatiales, temporelles et spatio-temporelles. Le second niveau apporte des détails aux concepts primaires, mais adaptés aux domaines d'application. Par exemple, la classe "SpatioTemporal-Primitive"

---

2. <http://www.hermit-reasoner.com/>

3. <http://owl.man.ac.uk/factplusplus/>

4. <https://github.com/stardog-union/pellet>



est précisée notamment par les concepts de mise en position des pièces dans un assemblage (soudage, clinchage, brasage, rivetage, etc.). Le troisième niveau correspond aux instances traitées par cette ontologie. Ce développement présente une description physique du produit. La description géométrique fait appel à des notions abstraites, comme par exemple celle de squelette d'assemblage, utilisée lors de la modélisation géométrique du produit par des outils de CAO. Cette ontologie ne comprend pas d'autres aspects numériques. Toutefois, elle est suffisamment générale pour être étendue et spécialisée, en ajoutant une description numérique du produit en plus de la description physique.

[Lu 15] proposent une ontologie enrichissant les échanges de données de produits couramment assurée par la norme STEP. Cette ontologie se concentre sur la gestion des données géométriques variables. Ces données décrivent les aspects paramétriques d'un modèle CAO liant des entités géométriques par rapport à des entités de références. Ce développement renforce la description de concept de modèle paramétré (qui peut être associé au concept de facteur d'un PEN), ainsi que du domaine de la représentation numérique de la géométrie d'un produit.

Faisant suite au projet européen PROMISE FP6 (PROduct lifecycle Management and Information tracking using Smart Embedded systems), dont le but était de fournir une infrastructure sécurisée pour l'échange et la gestion des données de produits durant toutes les phases du cycle de vie, l'ontologie SOM (Semantic Object Model) a été proposée [Matsokis 10]. Cette ontologie repose sur un modèle de donnée originel structurant les données de définition du produit en rapport avec les différentes phases du cycle de vie. Les données prises en compte décrivent les activités, les ressources (documents, matériels, humaines), la conception du produit (modèle CAO, nomenclatures, coûts, essais, etc.) ou encore le produit physique. SOM ne dispose pas de description spécifique pour les processus de simulation numérique ou de PEN. Toutefois, le processus de simulation numérique reposant sur le modèle CAO du produit, ce lien entre le produit et le modèle CAO offre une possibilité particulière d'étendre cette ontologie.

ONTO-PDM est une ontologie produit gérant les données hétérogènes, telles que celles décrivant les matériaux utilisés dans le produit, les relations entre les composants et les produits, les versions, les outils de fabrication, etc. [Panetto 12]. Les normes ISO 10303 et IEC 62264 ont servi de base pour le développement de cette ontologie. La norme ISO 10303 apporte une description partageable des données du produit (géométrie, données d'identification, etc.). La norme IEC 62264 assure le lien entre l'ontologie et les systèmes MES (Manufacturing Execution System) et ERP (Enterprise Resource Planning). ONTO-PDM ne semble pas supporter en l'état des données relatives aux processus de simulation et de PEN, mais l'utilisation de l'ISO 10303 comme base de ONTO-PDM permet une compatibilité direct avec les AP décrivant les données plus spécifiques, telles que l'AP 209.

PARO [Le Duigou 11] est à l'origine un modèle de données ayant évolué en une description ontologique des données relatives aux produits, aux activités, aux ressources et aux organisations. Ces quatre concepts forment le cœur générique de PARO et sont réutilisables notamment pour le processus de [PEN](#). Le concept de processus est décomposé selon la phase du cycle de vie. Le processus de [PEN](#) faisant partie du processus de conception, il se place naturellement comme une sous-classe de la classe "DesignProcess", elle-même sous-classe de "BOLProcess", regroupant les processus du début du cycle de vie du produit. Cette ontologie représente également les concepts de projet, de version, de composant et d'assemblage, de ressources matérielles et logiciels ou encore de documents.

OntoStep [Fiorentini 07, Barbau 12] est une ontologie basée sur les modèles de données CPM, OAM ainsi que sur les modèles spécifiés par les AP203, 214 et 239 de la norme [ISO 10303 STEP](#) (sous-section [2.2.1](#)). Ces modèles ont été combinés et enrichis avec des capacités sémantiques et de raisonnements. OntoStep couvre ainsi les données géométriques, les données relatives aux fonctions, aux exigences, aux comportements et aux choix de conception du produit. Cette ontologie ne dispose d'aucune spécialisation concernant les activités d'analyses.

Certaines ontologies ont été développées plus spécifiquement pour le processus de fabrication, comme par exemple OntoSTEP-NC [Danjou 14] ou InVor [Rehage 15]. OntoSTEP-NC, basé sur les standards [ISO 10303 AP238](#) et [ISO 10649](#), permet d'assurer une interopérabilité au sein des chaînes numériques d'industrialisation, comprenant les systèmes [PDM](#), ERP, MES et MPM. Les canaux d'échanges de connaissances offerts par cette ontologie facilitent la réutilisation des connaissances pour améliorer le processus de fabrication. InVor est une ontologie dédiée aux processus de fabrication, et plus spécifiquement aux processus d'usinage. L'utilisation de cette ontologie vise à améliorer la flexibilité des moyens de fabrication. En cas de panne ou d'utilisation d'une machine, une autre machine peut être sélectionnée rapidement selon ses capacités sans changer le programme d'usinage.

Concernant la simulation numérique, [Grosse 05] ont proposé une ontologie fournissant une représentation générique des connaissances liées au processus de simulation numérique. Cette ontologie décrit les modèles numériques, les simulations, les types d'analyses, les ressources matérielles et logiciels, les justifications des idéalizations. Les liens avec le produit et les processus sont également assurés. Cette description concorde avec de nombreux concepts nécessaires aux processus de [PEN](#). Toutefois, cette ontologie ne décrit pas le processus de [PEN](#) en tant que tel.

Une autre ontologie, proposée par [Assouroko 14], couvre les aspects de définition du produit (fonctions, exigences, représentation géométrique, comportement, etc.) et de simulation du comportement du produit (concepts de modèle numérique, de boucle de simulations, de résultat, de vérification des exigences, etc.). Cette ontologie supporte le

lien entre la représentation géométrique CAO et les activités de simulations, par le concept générique d'entité (ex : entité mécanique, entité de simulation, entité d'exigence.). Cette ontologie couvre les concepts nécessaires à la mise en place d'un processus de PEN, mais, là encore, sans prendre en compte le concept même de processus de PEN.

### Ontologies spécifiques

Cette seconde partie présente les ontologies spécialisées supportant des données spécifiques aux processus de PEN. La recherche des travaux existants a été effectuée sans prendre en compte le critère de domaine d'application. Le but de cette démarche est d'identifier les ontologies spécialisées pour les réutiliser dans le domaine de la mécanique, afin de favoriser l'échange des connaissances en milieu multi-disciplinaire.

EXPO [Soldatova 06] couvre le domaine sémantique des expériences scientifiques. Cette ontologie décrit, entre autres, les concepts d'expériences physiques et numériques, les résultats associés, les tests statistiques (tests d'hypothèses), la notion de facteur et de variable d'intérêt. Certains concepts permettant d'assurer une traçabilité minimale des expériences sont intégrés, tels que le titre ou le nom de l'auteur. EXPO est basé sur une méta-ontologie, plus générale, nommée SUMO (Suggested Upper Merged Ontology) [Pease 02], couvrant les concepts de temps, d'unité physique, de processus, d'objet, etc.

Basée sur EXPO, "Exposé" [Vanschoren 10] décrit le domaine plus spécifique des méthodes d'analyse de données ("data mining"). "Exposé" prend notamment en compte les concepts de données, d'ensemble de données (pour l'apprentissage et la validation), de paramètres (à la fois pour les modèles et les algorithmes générant ces modèles), de prédictivité, de ressources de calcul, de modèles de régression. Parmi les modèles de régression sont décrits les modèles de réseau de neurones, de processus gaussien (menant au krigeage), d'arbre de décisions, etc. Exposé définit également le concept de plan d'expériences et décrit les processus expérimentaux. Exposé détaille de nombreux concepts impliqués dans un processus de PEN. Elle décrit de façon exhaustive toutes les notions du domaine de l'analyse de données. En revanche, elle ne montre aucune trace d'intégration à un secteur particulier ou avec une stratégie PLM. Elle ne détaille pas les différents types de PEN, les algorithmes d'optimisations ni les méthodes d'analyse de sensibilité par exemple. Exposé n'est pas centrée sur le domaine des PEN.

Enfin, STATO<sup>5</sup> est une ontologie dédiée aux domaines de la biologie, de la génomique et de la neurologie. Certaines classes comprennent de nombreux concepts telles que les tests d'hypothèses, les mesures, les lois de probabilités, les modèles de régression, les variables et les plans d'expériences physiques. Ces concepts permettent de définir les facteurs d'un

---

5. <http://stato-ontology.org/>

PEN, certaines méthodes d'analyse de la variance pour l'analyse de sensibilité, ainsi que des méthodes de production d'un modèle de substitution.

### 2.2.3 Synthèse sur les modèles de connaissances

La première partie de l'état de l'art a permis d'identifier les domaines sémantiques et les concepts actuellement couverts par les développements existants. De nombreux concepts utilisés dans le processus de PEN sont déjà décrits dans les modèles de données et les ontologies. La figure 2.10 synthétise cet état de l'art en catégorisant les différentes ontologies et les modèles de données selon leurs capacités à supporter les connaissances relatives aux produits ou au processus, et selon les activités pour lesquels ils ont été développés. Les activités prises en compte sont celles du début de cycle de vie du produit, à savoir la conception préliminaire, la conception détaillée (simulation numérique) et la fabrication. A celles-ci s'ajoutent spécifiquement le processus de PEN et les domaines détachés du développement de produits dans lesquels les PEN ont une utilité.

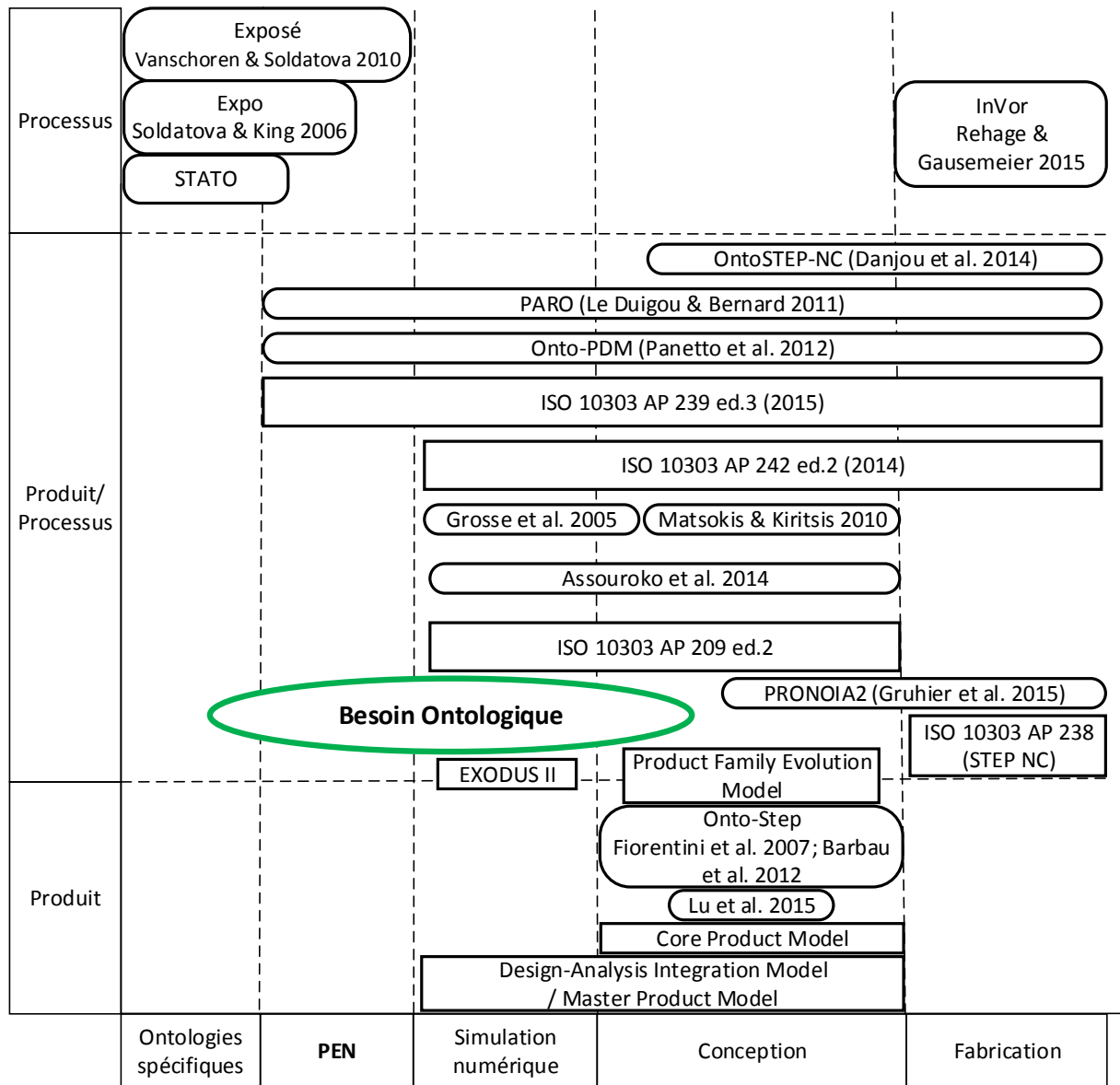
Cette représentation synthétique permet d'identifier le manque en termes de description ontologique. Il n'existe pas, selon cet état de l'art, d'ontologie ou de modèles de données centrés sur le processus de PEN assurant la connexion avec les connaissances portant sur le produit et sur les processus déployés durant le développement d'un produit. En l'état actuel, l'utilisation de ces ontologies ne permet pas d'assurer une capitalisation des connaissances communément comprises et partageables, ni une traçabilité des connaissances et des données durant le cycle de vie du produit, en intégrant le processus de PEN dans le processus de développement.

Cet état de l'art a permis également de repérer globalement les concepts à relier lors de la conception de l'ontologie qui sera proposée en réponse à ce manque. Ces concepts seront détaillés plus précisément au cours du chapitre 3.

Une telle ontologie contribuera à structurer et partager les connaissances tout au long du cycle de vie du produit et à travers une entreprise étendue. Cette structuration amène à une capitalisation des connaissances. Ces connaissances, en plus d'assurer une traçabilité, peuvent être ré-utiliser lors de processus de décisions.

La section suivante a pour objet la réutilisation des connaissances capitalisées, pour diriger les concepteurs vers une configuration de processus de PEN efficace et adaptée à leurs besoins et accélérer l'étape de configuration d'un processus de PEN.

FIGURE 2.10 – Classification synthétique des ontologies étudiées



**Besoin Ontologique**

Modèles de données      Ontologies

## 2.3 Réutilisation des connaissances

Cette troisième section de l'état de l'art traite des méthodes pour réutiliser les connaissances capitalisées dans l'entreprise. L'objectif est d'identifier les méthodes permettant d'assister le concepteur dans ses choix.

Un développement notable par rapport au contexte des processus de **PEN** est le système proposée par [Crombecq 11, Gorissen 09], nommé SUMO (SURrogate MOdeling)<sup>6</sup>. SUMO rejoint l'objectif de ces recherches, en sélectionnant automatiquement le type de modèle empirique et le type de métaheuristique pour un **PEN** adaptatif. L'objectif de ce système est de choisir l'ensemble des méthodes et paramètres pour l'obtention du modèle de substitution le plus fidèle au modèle numérique simulant le comportement du produit. Ce système teste toutes les combinaisons possibles afin d'identifier le meilleur modèle de substitution. Il enrichit le **PEN** initial si les exigences de prédictivité du modèle de substitution ne sont pas atteintes. Cette application ne couvre donc pas l'ensemble des choix couverts dans ces travaux (type de **PEN**, nombre d'expériences initial, etc.) mais, par cette orientation, elle limite le nombre d'expériences numériques nécessaires.

Un autre apport pour l'aide à la décision dans un contexte de simulation numérique a été proposé par [Kestel 16]. Ce système fournit une assistance au concepteur pour l'élaboration de modèles de simulation par éléments finis. Il se base sur les données et les connaissances capitalisées, en utilisant des méthodes d'analyses propres au domaine de l'IA (text mining, data mining).

Notre objectif est de traiter l'ensemble des choix possibles, pour tous les types de méthodes liées aux processus de **PEN**. Le système SUMO n'est donc pas applicable dans ce contexte. Faire varier le type de **PEN**, par exemple, nécessiterait d'exécuter de nouvelles expériences numériques à chaque essai. Une telle stratégie aurait ainsi un coût et/ou un temps d'exécution prohibitif (selon les ressources informatiques disponibles et la complexité du modèle numérique).

La section précédente de l'état de l'art a révélé un manque en termes de gestion et de structuration des connaissances spécifiques aux **PEN**. Une description ontologique de ce domaine apporte une base pour la réutilisation des connaissances capitalisées. Cette section de l'état de l'art se concentre sur les méthodes de réutilisation des connaissances, exploitant directement les connaissances exprimées par les experts ou exploitant les données capitalisées, pour créer de nouvelles connaissances et fournir une aide à la décision. Le but de cette démarche de recherche est de sélectionner le type de méthode de réutilisation des connaissances le plus adapté à notre contexte.

Cette section détaille au préalable les critères de sélection de la méthode de réutilisation

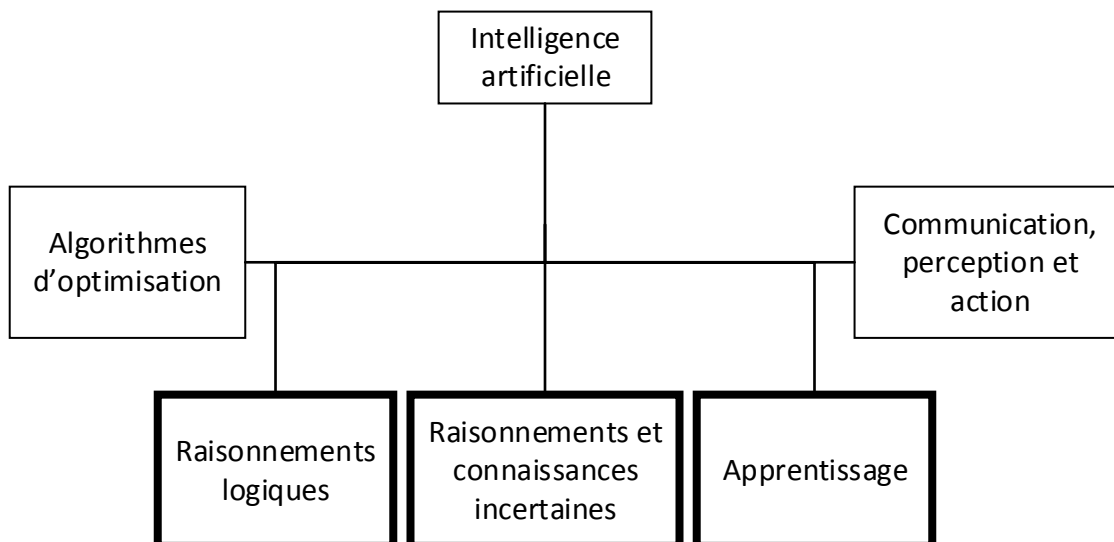
---

6. <http://www.sumo.intec.ugent.be>.

des connaissances, puis passe en revue les méthodes principales utilisées en IA. La figure 2.11 présente les grands domaines de l'IA relevés dans l'ouvrage de [Russel 10] :

- le domaine de la résolution de problème, lié aux méthodes d'optimisations et à la théorie des jeux ;
- le domaine du raisonnement logique, de la connaissance et de la planification ;
- le domaine du raisonnement basé sur des connaissances incertaines, utilisant des méthodes probabilistes ;
- le domaine de l'apprentissage (Machine Learning) permettant d'actualiser ou de créer des connaissances à partir de nouvelles expériences ;
- le domaine de la perception, de la communication et de la robotique, mobilisant des méthodes de reconnaissances du langage et des images.

FIGURE 2.11 – Principaux domaines de l'intelligence artificielle, issue de [Russel 10]



En gras, les domaines pris en compte dans l'état de l'art.

Ces recherches se sont concentrées sur les méthodes de raisonnement basées sur la connaissance et l'apprentissage. Les domaines présentés ici ne sont pas cloisonnés et certaines méthodes étudiées appartiennent à plusieurs domaines. Ces méthodes sont comparées par rapport aux critères énoncés. Une méthode est alors sélectionnée, afin de mettre en place un système à base de connaissances.

### 2.3.1 Critères d'analyses

Pour assurer une aide à la décision la plus efficace possible, il est nécessaire d'identifier une solution assurant les fonctions détaillées ci-dessous.

1. **Un raisonnement explicite.** L'objectif de ce système est de fournir une aide à la décision et ne prétend pas remplacer l'utilisateur. Le moteur d'inférences ne doit pas se comporter comme une boîte noire. Les propositions du système doivent être accompagnées d'explications compréhensibles permettant une appropriation de ces conclusions et la création de nouvelles connaissances.
2. **Une capacité d'apprentissage.** Le système doit pouvoir supporter l'enrichissement progressif de la base de connaissances. Les connaissances existantes concernant la configuration du processus de PEN sont parfois floues ou tacites. Une capacité d'apprentissage automatique à partir des données capitalisées permettra d'enrichir le raisonnement et le faire évoluer au cours du temps. Dans un but de découverte de nouvelles connaissances, il est nécessaire de se baser sur les données récoltées et de réaliser un raisonnement empirique, au moins en partie.
3. **Le support des connaissances expertes.** Les connaissances explicites, même incertaines, sont essentielles et ne peuvent pas être ignorées. Les connaissances expertes doivent être intégrées au sein du moteur d'inférences. Toutefois, le moteur d'inférences doit pouvoir se passer de telles connaissances si elles sont manquantes.
4. **Le support des données discrètes et continues.** Comme identifié dans le chapitre 1, la majorité des données définissant une configuration sont des données discrètes. Par exemple, le choix d'un type de modèle empirique se fait au sein d'un ensemble fini et discret de méthodes existantes  $E = [\text{Réseaux de neurones, Polynômes, Krigeage, etc.}]$ . D'autres données peuvent être continues, comme certains paramètres de certaines méthodes (ex : seuil de convergence d'un algorithme d'optimisation) ou les valeurs des indicateurs de performance d'une configuration donnée. Le moteur d'inférence doit pouvoir supporter les deux types de données.
5. **Le support des méthodes à comportement aléatoire.** Certaines méthodes utilisées lors d'un processus de PEN peuvent avoir un comportement aléatoire et donc difficilement prévisible. On peut citer par exemple les algorithmes évolutionnaires, utilisés notamment pour un enrichissement adaptatif d'un PEN. La maîtrise de ce type de méthode est liée à des connaissances tacites.
6. **Une performance prédictive suffisante.** Le moteur d'inférences doit, d'une part, proposer des configurations dans un temps raisonnable, et d'autre part, faire des prédictions correctes de configuration.

Une revue des développements concernant les systèmes experts [Shu-Hsien Liao 05, Sahin 12] ont permis d'identifier les principales méthodes existantes : les systèmes experts,



les systèmes à base de réseaux de neurones, les réseaux bayésiens, le raisonnement à base de cas, les arbres de décisions et les systèmes hybrides. Dans la suite, chacune de ces méthodes est analysée suivant les critères numérotés définis précédemment.

### 2.3.2 Méthodes de réutilisation

#### Les systèmes experts

Les systèmes experts rassemblent une base de règles logiques sur lesquelles sont effectués des raisonnements. Le début des années 1990 ont vu l'émergence de plusieurs développements sous la forme de systèmes experts pour la mise en place d'une démarche de plans d'expériences.

DEXPERT [Lorenzen 92] est un système expert, dont le raisonnement est basé sur un ensemble pré-défini de règles logiques. L'objectif de ce système expert est de guider l'utilisateur dans sa démarche expérimentale. Il présente une interface homme/machine adaptable suivant le profil de l'utilisateur, permettant de proposer des plans d'expériences plus efficaces pour un objectif donné tout en expliquant son raisonnement. DEXPERT couvre une partie des concepts déjà détaillés précédemment. Les objectifs d'analyse de sensibilité et de robustesse, d'optimisation et de production de modèles empiriques sont considérés. Toutefois, les méthodes utilisées pour atteindre chacun de ces objectifs sont peu nombreuses. Par exemple :

- les types de plan d'expériences sont limités aux types classiques (plan factoriels, plans de Plackett-Burman, plans de Box-Behnken, etc.) ;
- seuls les modèles empiriques polynômiaux sont considérés, et seule l'erreur de précision est calculée ;
- l'optimisation est réalisée par une heuristique simple pour repérer le minimum de la fonction polynômiale créée.

DEXPERT se base essentiellement sur les modèles polynomiaux estimés pour comparer les plans d'expériences entre eux. Ce système peut proposer des pistes d'amélioration du plan, en modifiant par exemple, son type, le nombre de facteurs pris en compte ou la définition de ces facteurs. La base de connaissances structure les données suivant une architecture orientée objet. Les évolutions d'un plan réalisées au cours de l'utilisation de DEXPERT sont représentées par un arbre, permettant ainsi une traçabilité des décisions. Ces données restent locales et ne sont pas partageables entre différents utilisateurs. Cette base de connaissances est utilisée essentiellement pour représenter les données. Ce système expert répond à la même problématique que celle de ces travaux de recherche. Ce développement reste limité au contexte des plans d'expériences physiques, présentant peu de choix

de méthodes, dans un cadre non-collaboratif. Il ne présente pas de capacité d'évolution automatique et ne peut se baser que sur des connaissances certaines et explicites pouvant être formalisées par des propositions logiques.

Un système similaire a été proposée par [Chen 91], guidant l'utilisateur dans un processus prédéfini de plan d'expériences. Ce système expert présente les mêmes limitations que DEXPERT. Les systèmes experts proposés par [Shraim 89] et [Weiner 92] suivent la même voie et présentent, eux aussi, les mêmes limitations que DEXPERT.

Une évolution de ce type de système est l'utilisation de règles de logiques floues. [Urrea 15] propose un système expert flou pour la sélection de matériaux adaptés à un contexte donné. L'utilisation de la logique floue permettent de formaliser les connaissances incertaines.

Plus récemment, une capacité d'apprentissage automatique a été intégrée dans des systèmes experts appliqués au domaine de la mécanique. [Dolšak 02] propose le système expert FEMDES. Ce système répond à une problématique similaire à celle considérée dans ces travaux de recherches, mais appliquée au processus de simulation numérique. Dans le cadre de la modélisation par éléments finis, la mise au point du maillage nécessite plusieurs essais, et donc plusieurs exécutions potentiellement coûteuses du modèle, pour enfin aboutir à un maillage offrant le meilleur compromis entre coût et précision des résultats. Pour réduire ce nombre d'essai, FEMDES apporte une aide à la conception du maillage. La base de règles est constituée de règles induites à partir d'une base de 10 maillages de référence issus de la littérature, et de certaines règles ajoutées manuellement. L'apprentissage est assuré par la méthode de "Inductive Logic Programming", permettant de déduire des règles logiques à partir d'exemples. FEMDES satisfait la plupart des critères énoncés précédemment, à savoir : une capacité d'apprentissage, une capacité d'enrichissement par l'intégration de connaissances expertes, une capacité pour expliquer le raisonnement et le support des données discrètes. La principale limitation vient de la sensibilité du système par rapport à la base d'apprentissage. Les exemples doivent être validés et cohérents entre-eux. La représentation de phénomènes aléatoires, comme on peut en trouver dans l'utilisation de métaheuristiques, est donc limitée. De plus, la formalisation d'une règle à partir d'un nombre restreint de cas limite la validité de ces règles au contexte spécifique de ces cas.

[Naranje 14] propose le système à base de connaissances INTDDD (INTElligent Design of Deep Drawing Die) pour une mise en place automatisée du processus d'emboutissage. Ce système expert permet, notamment, la conception automatique des moules. Le raisonnement s'appuie sur une base de plus de 1000 règles pré-définies.

L'inconvénient majeur des systèmes experts est qu'ils reposent sur l'hypothèse de connaissances parfaitement et totalement acquises et retranscrites. Sachant que certaines connaissances sont tacites, cette hypothèse n'est pas toujours vérifiée. L'ajout d'une

capacité d'apprentissage permet d'enrichir la base de connaissance, mais les données d'apprentissage doivent être validées et cohérentes entre elles, ce qui limite le support de comportements aléatoires. Par exemple, un processus de [PEN](#) adaptatif basé sur une métaheuristique aura des performances non-répétables. Plusieurs exécutions du même processus peuvent engendrer des résultats très différents, ce qui peut aboutir à la création de règles contradictoires. Enfin, un système expert ne peut pas prédire, en lui-même, la valeur d'une variable continue. Ainsi, généralement, les systèmes experts ne respectent pas les critères n°3, n°4 et n°5.

### Les réseaux de neurones

Les réseaux de neurones, déjà introduits au chapitre 1, sont aussi utilisés pour fournir une aide à la décision. Dans le cadre de ces travaux, seuls les réseaux de neurones de type "perceptron multi-couches" (PMC) sont considérés. D'autres types de réseau existent, comme les réseaux convolutionnels, les cartes auto-organisatrices ou encore les réseaux récurrents [Ding 09]. Ils sont à la base d'une branche de l'[IA](#), dite connexionniste.

Les PMC ont été utilisés, par exemple, dans le cadre de la prédiction de performance de moteurs diesel [Tasdemir 11]. Les auteurs comparent les performances d'un PMC avec celui d'un système expert s'appuyant sur un ensemble de règles floues.

Plus exactement, le réseau utilisé est de type "back-propagation multi-layer feed forward network", ce qui signifie que son apprentissage est assuré par un algorithme de rétro-propagation et qu'il est constitué de plusieurs couches cachées de neurones. Outre le fait que l'application de ces deux méthodes aboutisse à des résultats très proches en terme de prédiction des performances du moteur, les auteurs relèvent l'avantage du réseau de neurones, à savoir sa capacité d'apprentissage. Les règles de logiques floues ont été déterminées à l'aide d'un expert au cours de réunions, alors que le réseau de neurones n'a eu besoin que d'analyser les données expérimentales à disposition.

[Cui 16b] ont utilisé les réseaux de neurones afin de fournir une aide au choix du type de modèle de régression pour produire un modèle de substitution. Ces travaux montrent que les réseaux de neurones peuvent être utilisés pour l'aide à la configuration de tout le processus de [PEN](#).

Bien que simple d'utilisation et performant [Naïm 07, Cui 16b], les réseaux de neurones ne respectent pas le critère d'explicabilité des raisonnements. Ils ne permettent ni de représenter clairement les connaissances ni d'identifier facilement les causes des résultats précis. Les réseaux de neurones sont couramment qualifiés de boîtes noires [Naïm 07] et ne respectent donc pas le critère n°1. Le critère n°3 n'est pas non plus respecté, puisque les réseaux de neurones ne peuvent pas intégrer des connaissances a priori.

## Les arbres de décisions

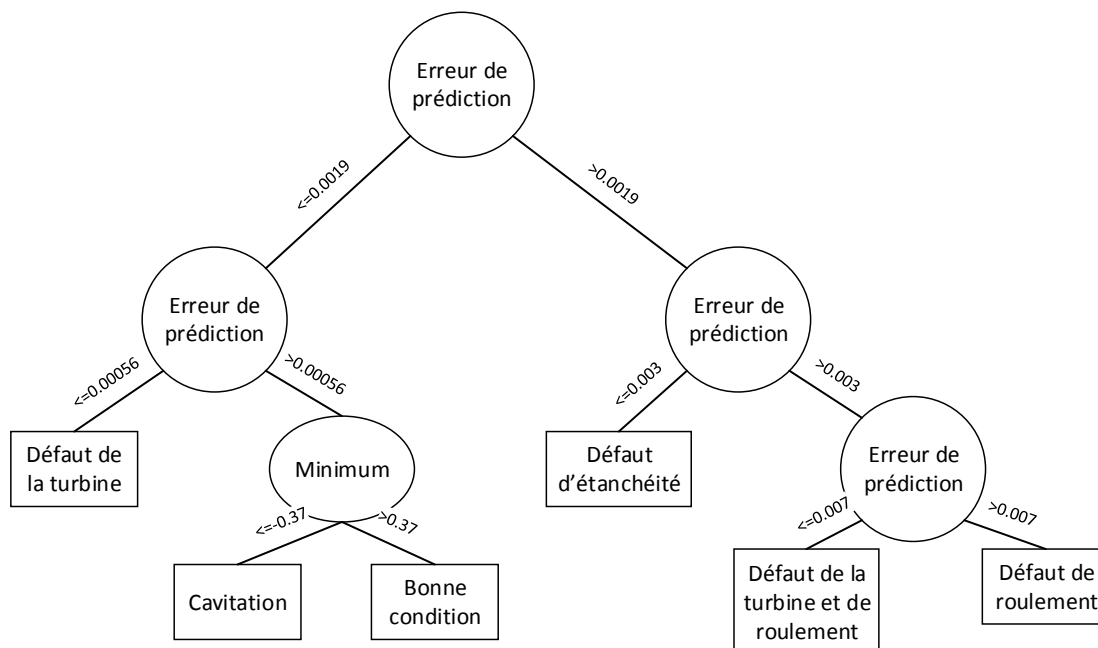
Les arbres de décisions sont généralement utilisés pour réaliser des classifications ou des régressions [Rakotomalala 05]. Ils reposent sur un partitionnement hiérarchisé des données d'apprentissage, représentable par un arbre. Il existe un certain nombre de types d'arbres de décision (CART [Breiman 01], C4.5 [Quinlan 96], CHAID [Kass 80]) ainsi que des évolutions, comme par exemple les forêts aléatoires [Breiman 01]. A chaque itération, une variable de segmentation est déterminée par un indicateur de qualité. De nombreux types existent. Pour la méthode CHAID, on utilise par exemple la statistique  $\chi^2$ , ou sa variante normalisée (le  $t$  de Tschuprow). La variable optimisant cet indicateur est choisie pour partitionner l'ensemble d'apprentissage pour l'itération en cours. Une variable de segmentation ne peut être utilisée que pour une seule itération. Cela génère une structure arborescente, composée d'une variable "racine", plusieurs niveaux de branches et enfin des feuilles, terminant l'arbre.

La construction d'un arbre de décision peut être instable si plusieurs variables ont un indicateur de qualité équivalent. Une petite modification de l'ensemble d'apprentissage peut entraîner une forte modification de la structure de l'arbre. On peut aussi constater une tendance au sur-apprentissage. Les "feuilles pures" terminant chaque branche de l'arbre peuvent ne pas avoir suffisamment d'individus pour être fiable lors de la prédiction. Cela tend à créer trop de catégories. Le contrôle de la taille de l'arbre s'avère déterminant pour garantir sa prédictivité. Des méthodes dites d'élagage permettent de supprimer des branches de l'arbre non-pertinentes. Les méthodes classiques (CHAID, CART, C4.5) ne traitent pas les interactions entre variables et nécessitent un ensemble d'apprentissage de grande taille. Les variables continues sont supportées mais uniquement par le biais d'une discrétisation. Les points de coupures pour les variables continues, seuils définissant deux ou plusieurs catégories pour cette variable, doivent être choisis a priori.

Les arbres de décisions ont, par exemple, été utilisés pour le diagnostic des défaillances de pompes centrifuges monobloc [Sakthivel 10]. L'arbre produit fournit une représentation explicite pour l'interprétation des indicateurs statistiques servant à suivre le comportement de la pompe (Figure 2.12).

Les arbres de décisions sont non-paramétriques (sans aucune hypothèse sur la distribution des données), robustes par rapport aux données aberrantes, capables de traiter tout type de données, et sont non-linéaires. De plus, ils sont facilement interprétables et lisibles. Les difficultés d'apprentissages peuvent être levées par des méthodes correctives, ou par l'utilisation de forêts aléatoires [Hastie 09]. Le seul inconvénient des arbres de décisions est l'impossibilité d'introduire des connaissances expertes. Ils ne respectent donc pas le critère n°3.

FIGURE 2.12 – Exemple d'arbre de décision, issue de [Sakthivel 10]



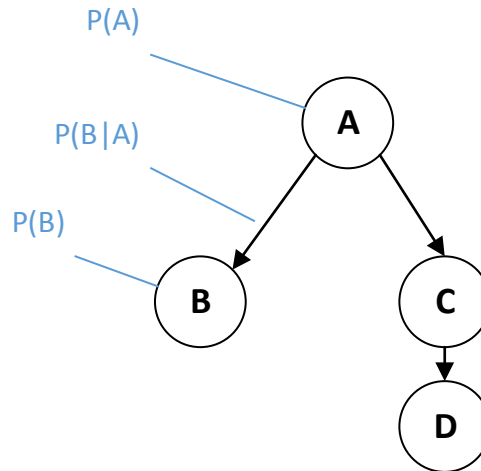
## Les réseaux bayésiens

Un réseau bayésien est représenté par un graphe acyclique orienté. Ce graphe est composé d'un ensemble de variables (liées à une loi de probabilité) et un ensemble d'arcs, représentant la dépendance entre les variables. Pour illustrer cette définition, le réseau bayésien de la Figure 2.13 est composé de 4 variables, A, B, C et D. A chaque variable est associée une loi de probabilité. Les arcs représentent les dépendances entre A et B, A et C, et C et D. Implicitement, B et D sont indépendants. L'orientation des arcs doit assurer l'absence de boucles pour que le réseau reste acyclique. L'orientation de l'arc A-B signifie que les états de A causent une modification des événements de B, et non l'inverse. Bien que la formule de Bayes permettent de retrouver les probabilités  $p(B|A)$ , cette orientation permet d'exprimer une causalité logique. Toutefois, le fait qu'un réseau bayésien soit causal n'est pas nécessaire pour obtenir de bonnes prédictions [Naïm 07].

L'ensemble des arcs orientés constitue la structure du réseau bayésien. L'ensemble des lois de probabilités conditionnelles, définie à partir de la structure, est appelé ensemble des paramètres du réseau bayésien. La structure et les paramètres peuvent être déterminés soit par l'expression d'une connaissance experte sous forme de probabilité, soit par apprentissage statistique à partir de données expérimentales. Ces deux approches peuvent être appliquées conjointement.

Il existe d'autres types de réseaux bayésiens, comme par exemple les réseaux bayésiens naïfs, les réseaux bayésiens naïfs augmentés par un arbre de décision, et l'approche "multi-net", prenant en compte la dépendance de la structure du réseau par rapport aux modalités

FIGURE 2.13 – Exemple de graphe d'un réseau bayésien



de chaque variable pour créer plusieurs sous-réseaux bayésiens simples à partir du réseau global [Naim 07].

Les réseaux bayésiens ont été utilisés par [Slanzi 14] pour définir automatiquement un plan d'expériences optimal pour des problèmes à grandes dimensions. [Hanafy 14] ont appliqué les réseaux bayésiens pour représenter les relations entre le produit et les systèmes de fabrication afin d'améliorer le taux d'utilisation des machines-outils. La géométrie du produit est caractérisée et le réseau bayésien permet d'assigner ce produit aux machines-outils adaptées automatiquement. La démarche inverse est également possible. Les réseaux bayésiens ont aussi été appliqués pour la prédiction de taux de défaillance des produits [Cai 11]. Les auteurs proposent un nouvel algorithme de construction de réseaux bayésiens, appelé "réseau bayésien conditionnel". [Correa 09] propose l'application des réseaux bayésiens pour la prédiction de la rugosité des surfaces après usinage.

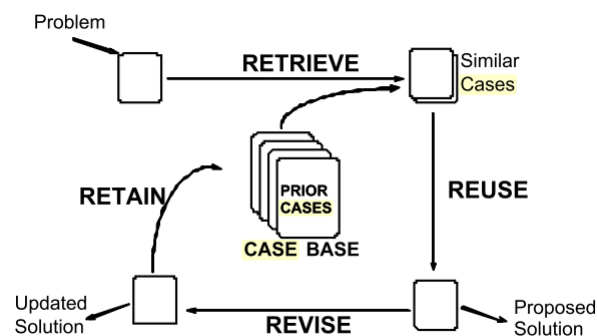
Les prédictions effectuées par un réseau bayésien peuvent être expliquées par sa représentation sous forme de graphe et par les lois de probabilités conditionnelles. Ils supportent l'introduction de connaissances expertes, en imposant des lois de probabilités, des dépendances ou des indépendances a priori. Ces connaissances ne sont pas obligatoires si des données sont disponibles, permettant de déterminer la structure et les paramètres par apprentissage. Ils peuvent supporter également les variables discrètes et continues, par la modélisation des lois de probabilités sous forme discrète ou continue. Enfin, les probabilités sont à même de représenter les connaissances incertaines et les comportements aléatoires. Par exemple, si pour une même configuration utilisant un algorithme génétique, 6 exécutions ont données de mauvais résultats et 4 de bons résultats, la probabilité que cette configuration donne de bons résultats est simplement de 0.4. Ainsi, les réseaux bayésiens satisfont tous les critères énoncés précédemment.

## Le raisonnement à base de cas

Le raisonnement à base de cas est une catégorie de méthode basée sur la réutilisation et l'adaptation de solutions pour un nouveau cas. Ce type de raisonnement se décompose en quatre étapes principales (Figure 2.14) [Aamodt 94] :

1. retrouver les cas les plus similaires. De nombreuses méthodes de mesure de similarité ont été proposées [Lopez de Mantaras 05] ;
2. réutiliser les connaissances de ces cas pour résoudre le problème considéré. La réutilisation peut être réalisée par l'adaptation du cas le plus similaire ou par la réutilisation de la méthode ayant abouti à ce cas ;
3. réviser la solution proposée. La solution issue des deux premières étapes n'étant qu'une proposition, elle peut être imparfaite et peut nécessiter des ajustements ;
4. retenir cette nouvelle solution. La base de cas est enrichie à chaque cycle.

FIGURE 2.14 – Principe de fonctionnement du raisonnement à base de cas. Illustration tirée de [Lopez de Mantaras 05], adaptées de [Aamodt 94]



Ce type de raisonnement a été appliqué, par exemple pour la simulation numérique. [Khan 14] propose un algorithme de raisonnement à base de cas pour une définition plus rapide des maillages par la méthode des éléments finis. [Liu 08] applique ce type de raisonnement pour faciliter le processus de conception de produit.

Le raisonnement à base de cas offre une capacité d'apprentissage et des explications claires. Il est toutefois obligatoire de spécifier des règles spécifiques au type de problème traité pour adapter les cas durant la phase de réutilisation. Le critère n°3, exigeant un support optionnel des connaissances, n'est donc pas respecté.

## Systemes hybrides

De nombreux travaux concernant les systèmes hybrides ont été menés au cours des 20 dernières années. Ce type de système combine plusieurs méthodes de raisonnement différents, créant ainsi un hybride, plus performant que les méthodes d'origine.

[Sahin 12] dressent une vue d'ensemble des productions scientifiques publiées dans 70 journaux et 16 conférences entre les années 1988 et 2010. Toutes ces productions scientifiques proposent des démarches hybrides, mêlant les systèmes à base de règles, les réseaux de neurones, les algorithmes évolutionnaires ou les arbres de décisions. Trois approches principales ont été identifiées, parmi les 45 systèmes hybrides relevés.

1. Les systèmes experts, à base de règles, intégrant un réseau de neurones. Par exemple, un tel système a été proposé pour la création automatique de modèles par éléments finis [Li 03] pour des fuselages d'avions.
2. Les réseaux de neurones flous. Ces systèmes combinent des règles floues et un réseau de neurones, comme par exemple ANFIS (adaptive-network-based fuzzy inference system) [Jang 93, Roy 13].
3. Les réseaux de neurones approximatifs, basé sur la théorie des ensembles approximatifs [Pawlak 94].

Par exemple, dans le but de prédire les résultats d'une simulation par la méthode des éléments finis, un système expert à base de règles floues a été développé [Subba Rao 07]. Les règles floues ont été déterminées à partir d'un échantillon de 275 cas, correspondant à la prise en compte des 11 possibilités par variables prises en compte. Les fonctions d'appartenance sont déterminées à partir de ces données et par un algorithme génétique.

Dans le cadre de ces travaux de recherches, les systèmes hybrides n'ont pas été, dans un premier temps, intégrés dans ce comparatif.



### 2.3.3 Comparaison des méthodes de réutilisation

Cet état de l'art sur les méthodes de réutilisation des connaissances a permis d'analyser les principales méthodes fournissant des prédictions de configurations de processus de **PEN**, et donc, apportant une aide aux concepteurs. Les réseaux bayésiens respectent tous les critères fixés préalablement (Table 2.2).

Toutefois, il est important de préciser que cet état de l'art n'est pas exhaustif. Les très nombreux développements réalisés sur les systèmes hybrides pourraient révéler des méthodes respectant elles-aussi ces critères.

Le critère n°6, portant sur les performances prédictive, n'a pas pu être évalué. Ce critère dépend du cas d'application, et les méthodes couvertes ne peuvent pas être comparées sur ce critère sans une application concrète.

TABLEAU 2.2 – Tableau comparatif des méthodes d'analyse des connaissances capitalisées

Type de raisonnement	Critères					
	Méthodes	N°1	N°2	N°3	N°4	N°5
Raisonnements logiques	Systèmes experts	✓	✓			
Apprentissage automatique	Réseaux de neurones		✓		✓	✓
	Arbres de décision	✓	✓		✓	✓
	Réseaux bayésiens	✓	✓	✓	✓	✓
Raisonnement à base de cas		✓	✓		✓	✓

Critères : 1-Explicabilité des décisions ; 2-Apprentissage ; 3-Support optionnel des connaissances expertes ; 4-Support des données discrètes et continues ; 5-Support des comportements aléatoires

## 2.4 Synthèse

La première partie de cet état de l'art s'est focalisée sur la gestion des connaissances. Les moyens existants en termes de modèles de données et de modèles de connaissances ontologiques ne couvrent pas spécifiquement les connaissances nécessaires à la configuration d'un processus de **PEN**. Certains concepts sont supportés par des modèles de données standardisés, notamment la norme **ISO 10303 STEP**, et par des ontologies. Ces ontologies décrivent soit des concepts généraux en lien avec le produit ou les processus en général, soit des concepts très précis, telles que les méthodes de régression ou la définition de variables aléatoires. Il est donc nécessaire de proposer une description ontologique spécifique pour le processus de **PEN**. Notre proposition de structuration ontologique sera composée de nouveaux concepts permettant de relier les développements ontologiques existants. L'application d'une telle ontologie permettra de capitaliser et de partager les connaissances relatives aux processus de **PEN**, en continuité de la gestion des connaissances relatives au cycle de vie du produit.

La seconde partie a analysé les méthodes de réutilisation de ces connaissances, pour aider le concepteur à configurer les processus de **PEN**. Les méthodes couvertes, appartenant au domaine de l'**IA**, comprennent les systèmes d'aide à la décision à base de règles, de réseaux de neurones, de réseaux bayésiens, d'arbres de décision et de raisonnement à base de cas. Les fonctions principales du système d'aide à la décision ont été définies et les méthodes couvertes par l'état de l'art ont été comparées en regard de ces fonctions. La méthode la plus appropriée, a priori, est celle des réseaux bayésiens.

Ces deux composantes peuvent être combinées pour former un système d'aide à la décision. Le chapitre suivant décrit la mise en œuvre du système d'aide à la décision, composée d'une ontologie spécifique au processus de **PEN** et d'un réseau bayésien permettant de proposer automatiquement de nouvelles configurations adaptées au problème de l'utilisateur. Le processus de **PEN** serait alors appliqué plus facilement tout en étant plus rapide et plus économe en expériences numériques. Les prises de décision pourront alors être plus rapides lors de la conception du produit, diminuant ainsi le temps de mise sur le marché du produit tout en renforçant sa qualité.

# Chapitre 3

## Proposition : système à base de connaissances pour l'aide à la configuration des processus de plan d'expériences numériques

■ *Quels que soient les progrès des connaissances humaines, il y aura toujours place pour l'ignorance et par suite pour le hasard et la probabilité.* ■

Emile Borel, *Le Hasard*, 1914, Librairie Félix Alcan, p12-13

### Sommaire

---

<b>3.1</b>	<b>Architecture globale du système</b>	<b>91</b>
<b>3.2</b>	<b>Ontologie du domaine des processus de plan d'expériences numériques</b>	<b>94</b>
3.2.1	Objectifs principaux	94
3.2.2	Description de l'ontologie pour les processus de plan d'expériences numériques	94
3.2.3	Retour sur les critères de conception de l'ontologie	100
3.2.4	Utilisation de l'ontologie	106
3.2.5	Apports de l'ontologie	109
<b>3.3</b>	<b>Réutilisation des connaissances capitalisées par réseaux bayésiens</b>	<b>110</b>
3.3.1	Conception du réseau bayésien	110
3.3.2	Validation des performances prédictives du réseau bayésien	119
3.3.3	Apports du réseau bayésien	124
<b>3.4</b>	<b>Synthèse et discussion</b>	<b>126</b>

---

*CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE  
À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES*

---

### CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

Ce chapitre décrit la solution proposée en réponse à la problématique formulée au chapitre 1, en se basant sur les conclusions de l'état de l'art, détaillés au chapitre 2. Cette solution prend la forme d'un système à base de connaissances remplissant deux fonctions principales :

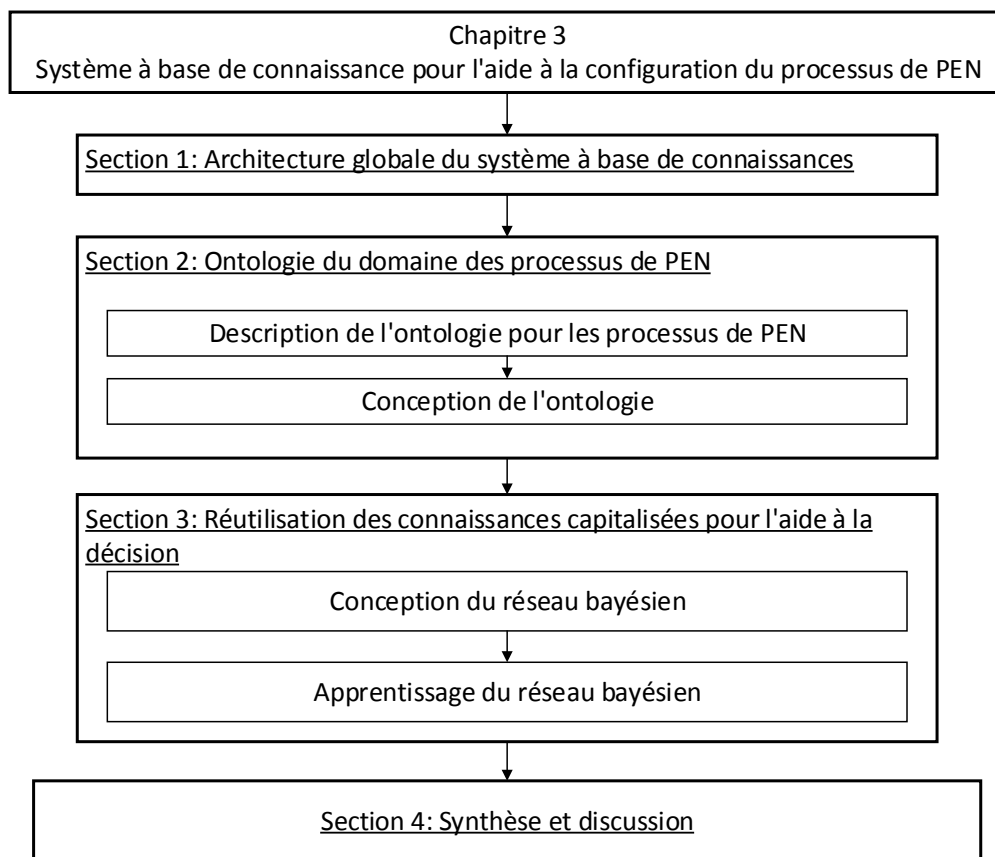
- la capitalisation structurée des connaissances spécifiques aux processus de **PEN** ;
- la réutilisation des connaissances capitalisées pour fournir une aide à la configuration du processus de **PEN**.

Ce chapitre se décompose en 4 sections (Figure 3.1) :

1. la section 1 présente l'architecture globale du système à base de connaissances proposé ;
2. la section 2 décrit l'ontologie proposée pour la gestion des connaissances relatives au processus de **PEN** ;
3. la section 3 présente le moteur d'inférences, basé sur les réseaux bayésiens, permettant d'analyser les connaissances disponibles afin de proposer de nouvelles configurations adaptées à de nouvelles situations ;
4. la section 4 synthétise la proposition et discute de ses avantages et de ses limites d'utilisation.

L'objectif de ce système à base de connaissances est de faciliter l'utilisation de processus de **PEN** au cours du développement du produit. Ce processus pourra être mis en place plus rapidement pour des modèles numériques complexes, et le produit pourra profiter d'une conception optimisée, plus robuste, plus fiable en amont du cycle de vie. L'entreprise pourra également profiter de ce système pour améliorer les processus collaboratifs au sein d'entreprises étendues, pour capitaliser les connaissances propres aux **PEN**, assurer la traçabilité de ses activités et de ses décisions et explorer de nouvelles solutions innovantes.

FIGURE 3.1 – Plan du chapitre 3

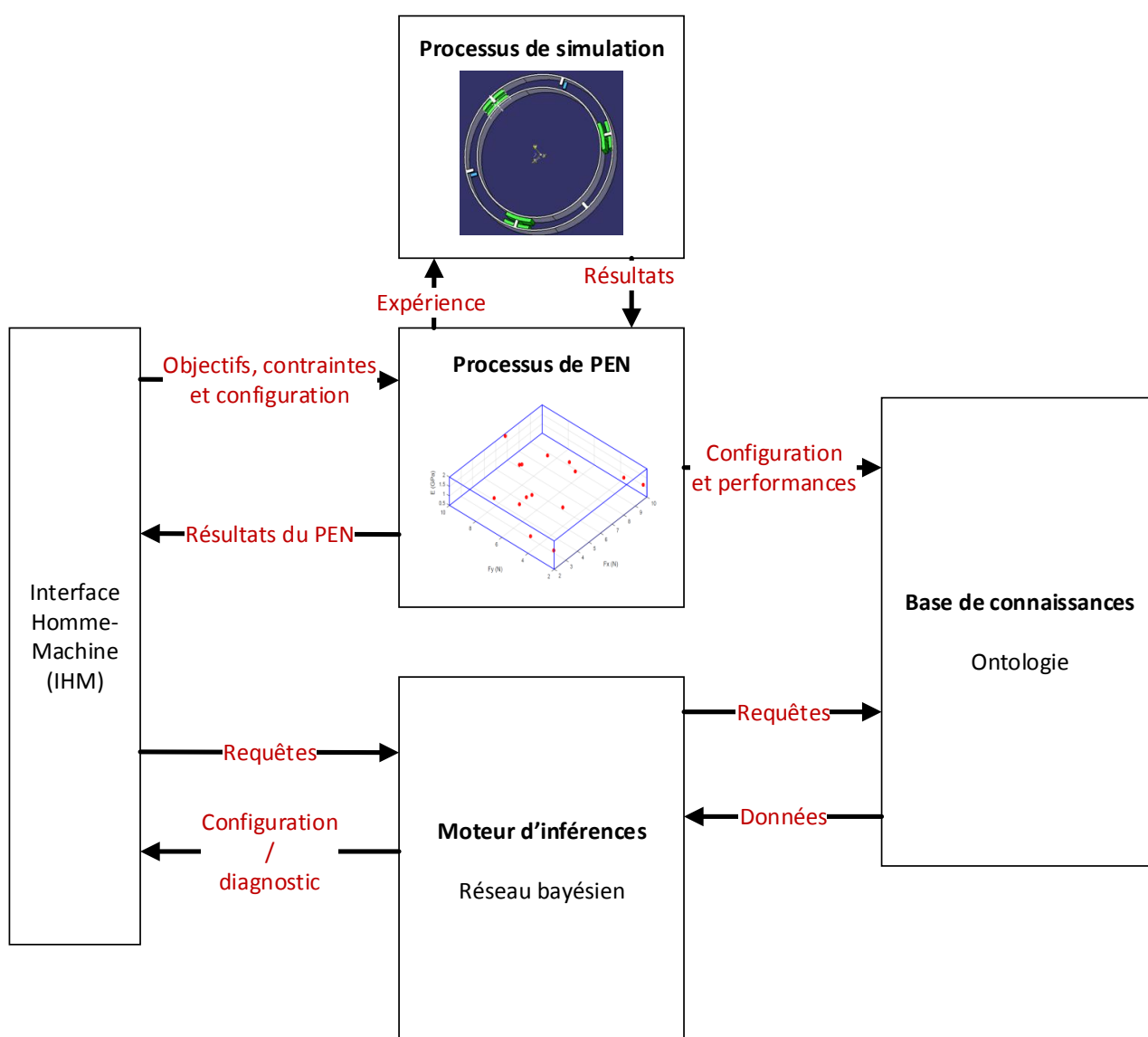


### 3.1 Architecture globale du système

Le système à base de connaissances proposé dans le cadre de ces travaux a pour objectif principal de réduire le temps nécessaire à la configuration du processus de PEN, notamment dans le cas où des méthodes complexes sont nécessaires pour réduire le coût d'exécution.

Ce système (Figure 3.2) est un outil d'aide à la décision. Il est centré autour de l'utilisateur, et n'a pas pour vocation à le remplacer.

FIGURE 3.2 – Architecture globale du système à base de connaissances proposé



Par le biais d'une interface, l'utilisateur spécifie un nouveau processus de **PEN**. Deux modes d'utilisation de ce système à base de connaissances sont à distinguer.

Dans le premier mode, l'utilisateur n'impose qu'une partie des éléments de la configuration, ou aucun, selon un contexte, un objectif et des contraintes données (budget, précision des résultats, ressources). Par exemple, il peut ne pas savoir combien d'expériences seront nécessaires pour obtenir un modèle de substitution précis et prédictif d'un modèle de simulation par éléments finis. Chaque expérience peut immobiliser les ressources de calcul de son équipe. S'il spécifie trop peu d'expériences, le modèle de substitution ne pourra pas se substituer au modèle initial et il sera nécessaire d'exécuter un nouveau **PEN** pour obtenir un modèle de substitution convenable. Si trop d'expériences sont spécifiées, des ressources auront pu être utilisées inutilement, les prises de décision en aval du processus de **PEN** seront plus tardives, et le modèle de substitution s'expose au sur-apprentissage, ce qui aboutira également à de mauvais résultats.

Dans le second mode d'utilisation, l'utilisateur propose un processus complet, mais ne connaît pas ses performances. En reprenant le même exemple, il souhaite savoir si, avec la configuration qu'il propose, le modèle de substitution sera à même de remplacer le modèle de simulation par éléments finis.

Pour ces deux modes, l'utilisateur pose le problème au système à base de connaissance. Suivant la situation initiale, les données issues de la base de connaissances sont extraites, en correspondance avec le problème auquel est confronté l'utilisateur. Les configurations de processus de **PEN** utilisées pour des cas similaires (même type de modèle de simulation, même objectif, même type de produit) sont extraites de la base de connaissances (structurée par l'ontologie), accompagnées de leurs performances associées. Par exemple, tous les processus de **PEN** utilisés pour une même gamme de moteur, par rapport à la simulation du comportement thermo-mécanique, dans l'objectif de produire un modèle de substitution, seront extraits. Tous les éléments constituant chaque processus de **PEN** pertinents, associés au temps d'exécution et aux valeurs des indicateurs de précision et de prédictivité du modèle de substitution sont récupérés.

Cet ensemble des données extraites est utilisé pour l'apprentissage du réseau bayésien. Un sous-ensemble des données extraites est utilisé pour l'apprentissage, et le reste permettra de tester ses performances prédictives. En effet, après l'apprentissage, le réseau bayésien prédira un résultat pour assister l'utilisateur dans ses choix. Dans le premier mode d'utilisation, le réseau bayésien prédira une configuration complète adaptée. Dans le second mode, il prédira la performance d'un processus de **PEN** complètement configuré. Il est alors obligatoire de quantifier le taux d'erreurs de prédiction du réseau bayésien. Comme pour les modèles de substitutions, des indicateurs de performance seront associés au réseau bayésien. Si les performances du processus de **PEN** définie par l'utilisateur ne sont pas



### CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

satisfaisantes, le système passera automatiquement dans le premier mode d'utilisation, en réutilisant le même réseau bayésien, pour proposer des alternatives.

Parmi les propositions faites par le réseau bayésien se trouveront les cas déjà traités, mais aussi des prédictions sur des configurations inédites. L'utilisateur pourra, au choix, rester sur une configuration connue, ou explorer de nouvelles solutions si elles s'avèrent plus performantes. L'utilisateur ne fait pas ce choix au hasard. Par le biais de l'interface, il a accès à toutes les informations lui permettant de connaître le risque qu'il prend en choisissant une nouvelle solution. Des mesures de probabilités que le réseau bayésien a d'avoir raison ou tort sont systématiquement associées aux prédictions. Suivant la gravité d'une erreur (une prédiction peut être fausse mais être à l'avantage de l'utilisateur, si par exemple une configuration de processus de PEN est plus performantes que prévue), l'utilisateur peut estimer le risque (probabilité \* gravité). Il est également libre de choisir une configuration proposée par le système et de la modifier, auquel cas le système pourra repasser dans le second cas d'utilisation.

Lorsque l'utilisateur a validé son choix, le processus de PEN est configuré (le type de plan, le nombre d'expériences, le type de modèle de substitution, etc.). Il est exécuté, les simulations sont lancées et les résultats du PEN et des différents analyses (production d'un modèle de substitution, analyse de sensibilité, etc.) sont récupérées. L'ensemble, constitué de la configuration et des résultats, est intégré dans la base de connaissance pour l'enrichir. Cet enrichissement est profitable qu'importe s'il s'agit d'un succès (économie de temps, excellentes performances du modèle de substitution, etc.) ou d'un échec.

Tous les essais des utilisateurs seront accessibles, grâce à l'ontologie, et les échecs et succès de chaque collaborateur viendront améliorer la productivité de toute l'équipe, ou de toutes les équipes.

Les deux sections suivantes détaillent respectivement l'architecture de l'ontologie, structurant la base de connaissances, et le fonctionnement du réseau bayésien.

## 3.2 Ontologie du domaine des processus de plan d'expériences numériques

Cette section présente l'ontologie proposée pour structurer la base de connaissances de l'entreprise. Cette ontologie va également servir de base pour le partage et la compréhension des connaissances capitalisées. Elle est conçue de façon à être compatible avec les ontologies existantes, pour les compléter avec une description des concepts spécifiques aux processus de **PEN**.

### 3.2.1 Objectifs principaux

L'ontologie proposée est définie pour combler le manque de description du domaine sémantique des **PEN**, constatés au cours de l'état de l'art (cf. chapitre 2). Cette ontologie répond aux objectifs suivants :

- supporter la capitalisation des connaissances relatives aux processus de **PEN**. Toutes les données relatives aux processus de **PEN** doivent être collectées et capitalisées avec une structuration sémantique commune ;
- partager les connaissances entre les différents utilisateurs. Les connaissances doivent être compréhensibles par toutes les personnes impliquées dans la réalisation et la gestion des processus de **PEN**. Cette compréhension doit être identique pour tous ;
- compléter et lier les modèles de données et de connaissances existants. L'état de l'art (cf. chapitre 2) a révélé l'existence d'ontologies de domaine voisins à celui des processus de **PEN**. La réutilisation est un des objectifs premier d'une ontologie, assurant la continuité des travaux en termes de formalisation de connaissances communément comprises.

### 3.2.2 Description de l'ontologie pour les processus de plan d'expériences numériques

Cette ontologie, nommée ODE (Ontology for numerical Design of Experiments) décrit les concepts essentiels spécifiques aux processus de **PEN**. ODE comporte, dans sa version actuelle, 77 classes, 31 propriétés liant ces classes (hors propriétés de sous-classe) et 38 types de données différents (Figure 3.3). Parmi ces 77 classes, trois classes principales se distinguent : la classe de processus de **PEN**, la classe de **PEN** et la classe de **PEN** adaptatif. Le langage utilisé est le langage OWL2-DL (cf. chapitre 2). La Figure 3.4 illustre les différentes classes définies sous Protégé<sup>1</sup>.

1. <http://protege.stanford.edu/products.php>

# CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

FIGURE 3.3 – Caractéristiques de ODE (avec instances)

Ontology metrics:	
<b>Metrics</b>	
Axiom	<b>24280</b>
Logical axiom count	<b>19469</b>
Declaration axioms count	<b>4798</b>
Class count	<b>77</b>
Object property count	<b>31</b>
Data property count	<b>38</b>
Individual count	<b>4644</b>
DL expressivity	ALCRQ(D)
<b>Class axioms</b>	
SubClassOf	<b>56</b>
EquivalentClasses	<b>5</b>
DisjointClasses	<b>1</b>
GCI count	0
Hidden GCI Count	<b>2</b>

FIGURE 3.4 – Arborescence de ODE

The screenshot displays the ODE ontology interface. The left pane shows a class hierarchy starting from owl:Thing, with 'Adaptive\_DoE' highlighted. The right pane shows 'Found 8 uses of Adaptive\_DoE' with a list of properties and their domains.

**Class hierarchy: Adaptive\_DoE**

- owl:Thing
  - STATO\_Statistical\_Methods
    - EXPO\_Factor ≡ DoE\_Factors
    - EXPO\_TargetVariable ≡ DoE\_Outputs
    - Expose\_predictive\_algorithm ≡ Surrogate\_ModelType
  - Infill\_Criterion
  - ISO10303\_AP209\_ed2
    - Finite\_elements\_Model
  - NDOE
    - Adaptive\_DoE**
    - NDoE\_Objective
      - Diagnostic
      - Exploration
      - Optimization
      - Sensitivity
      - Surrogate\_Modelling
    - NDoE\_Process
      - NDoE\_Process\_Results
        - NDoE\_Results
          - SurrogateModel\_Results
      - NDoE\_ProcessConstraints
      - NDoE\_Type ≡ STATO\_study\_design
        - Factorial
        - Space\_Filling\_Designs
    - NDoEModels
      - Initial\_Models
        - Surrogate\_Model
      - Optimization\_Algorithm
        - Gradient\_Based
        - NonGradient\_Based
    - PPRO\_Resources ≡ Resources
    - Quality\_Indicators
      - Accuracy
      - Predictivity
    - Resources ≡ PPRO\_Resources
      - Hardware
      - Software
    - Statistical\_Distribution
    - STATO\_study\_design ≡ NDoE\_Type
    - Surrogate\_ModelType ≡ Expose\_predictive\_algorithm
      - Artificial\_Neural\_Network
      - Kriging
      - Polynomial
    - Surrogate\_Validation\_Methods
      - Bootstrap
      - Leave\_One\_Out
    - Variables
      - Model\_Outputs
      - Model\_Parameters

**Usage: Adaptive\_DoE**

Found 8 uses of Adaptive\_DoE

- Adaptive\_DoE
  - Adaptive\_DoE SubClassOf NDOE
  - Class: Adaptive\_DoE
- AddedExp
  - AddedExp Domain Adaptive\_DoE
- AddedExp\_Limit
  - AddedExp\_Limit Domain Adaptive\_DoE
- hasInfill\_Criterion
  - hasInfill\_Criterion Domain Adaptive\_DoE
- hasOptim\_Algo
  - hasOptim\_Algo Domain Adaptive\_DoE
- hasQuality\_Objectives
  - hasQuality\_Objectives Domain Adaptive\_DoE
- improves\_Surrogate\_Model
  - improves\_Surrogate\_Model Domain Adaptive\_DoE

Dans la suite de la description de ODE, les entités propres à ODE sont écrites avec une police d'écriture spécifique pour les différencier des entités appartenant à d'autres ontologies. De plus, pour favoriser le partage de connaissances, ODE a été écrite en anglais. Par exemple, la classe représentant le concept de **PEN** sera noté "NDoE" (Numerical Design of Experiments).

Le concept principal, représenté par la classe "NDoE\_Process" est celui du processus de **PEN**. Il assure le lien entre le domaine des autres processus mis en place lors de la conception d'un produit et tous les concepts appartenant spécifiquement au processus de **PEN**.

Le processus de **PEN** (Figure 3.5) est défini par :

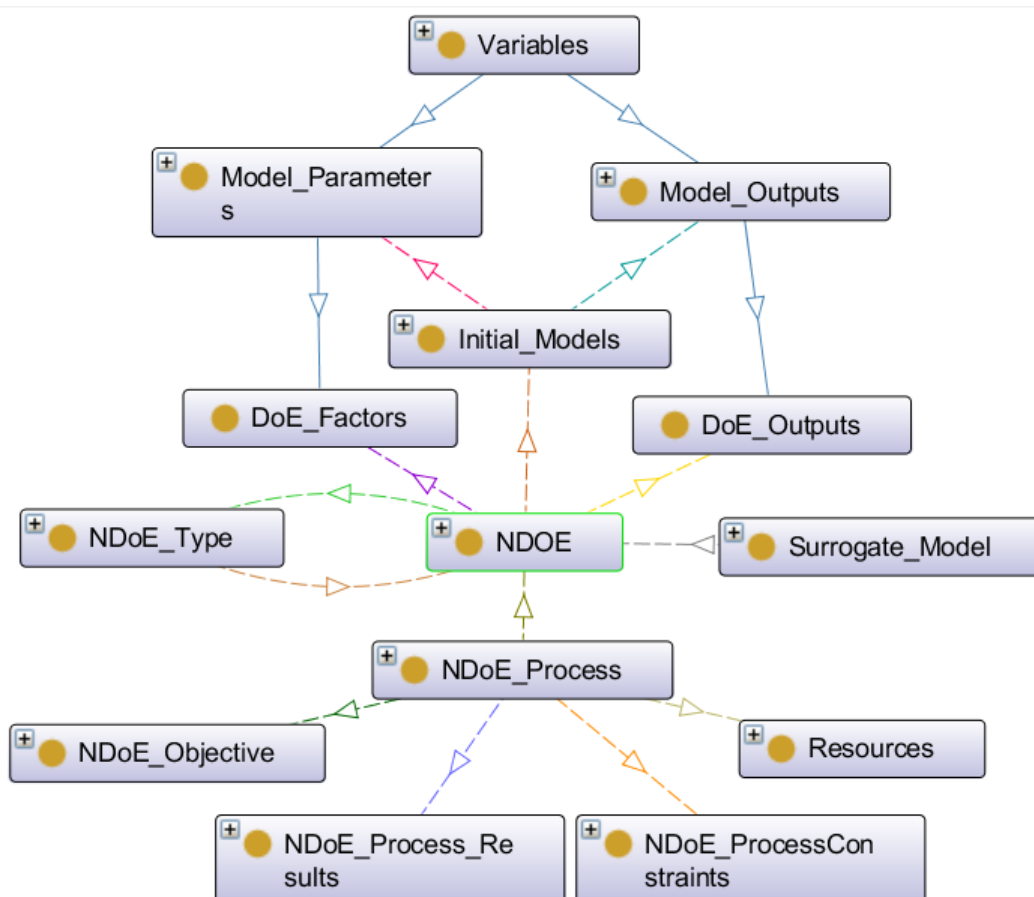
- un unique objectif (classe "NDoE\_Objective") parmi les 4 considérés (cf. chapitre 1).
- des contraintes, telles qu'une durée maximale ou un niveau de précision minimal pour un modèle de substitution (classe "NDoE\_ProcessConstraints");
- au moins un **PEN** (classe "NDoE");
- des résultats (classe "NDoE\_ProcessResults"), tels que les résultats de chaque expérience, les indices de sensibilité, le temps d'exécution, les modèles de substitution produits, etc. ;
- des données de contexte liées au projet. Elles ne sont pas représentées par des classes, mais par des données rattachées aux classes définies dans ODE. Elles couvrent l'identité de toutes les parties prenantes, les données liées au produit concerné, les données de traçabilité (dates, versions, configurations, etc.). Ces données comprennent également toutes les données du produit virtuel liées à l'étude en cours (type de modèle numérique, version du modèle, type de simulation, etc.);
- des ressources (classe "Resources"), telles que les logiciels et les moyens de calculs à disposition (par exemple, le nombre et la nature des processeurs disponibles, la quantité de mémoire vive disponible, etc.). Les données concernant les ressources déterminent le budget de calcul disponible et le temps d'exécution d'une simulation. Ces données peuvent aussi donner des informations sur la complexité de l'étude, dans le cas où plusieurs logiciels de simulations seraient couplés.

Le concept de **PEN** (classe "NDoE") rassemble quant à lui toutes les données spécifiques au processus de **PEN**. Il s'agit du concept permettant la définition de toute la configuration du processus. Ce concept (Figure 3.5) est défini par :

- exactement un type de plan. Le type conditionne la répartition des expériences dans le domaine expérimental (cf. chapitre 1) (classe "NDoE\_Type");
- un nombre d'expériences. Ce nombre, entier strictement positif, donne le nombre de simulations à réaliser ;

- exactement un modèle numérique initial (classe "Initial\_Models") pour l'exécution des expériences numériques. Dans le cas de simulations multi-physiques, le modèle numérique initial regroupe toutes les physiques étudiées. Le modèle est supposé fonctionnel et physiquement valide. Il peut s'agir d'un modèle par éléments finis par exemple, ou d'un modèle de substitution produit par un processus de PEN antérieur ;
- au moins un facteur (classe "DoE\_Factors"). Les facteurs sont issus de l'ensemble des paramètres du modèle numérique initial. Un facteur est défini à partir d'un paramètre du modèle numérique initial, par le PEN dans lequel il est impliqué et par une loi de probabilité (s'il s'agit d'un paramètre continu) ;
- au moins une variable d'intérêt (classe "DoE\_Outputs"), appartenant à l'ensemble des réponses du modèle numérique initial ;
- un modèle empirique (classe "Surrogate\_Model"), qui peut sous certaines conditions, servir de modèle de substitution.

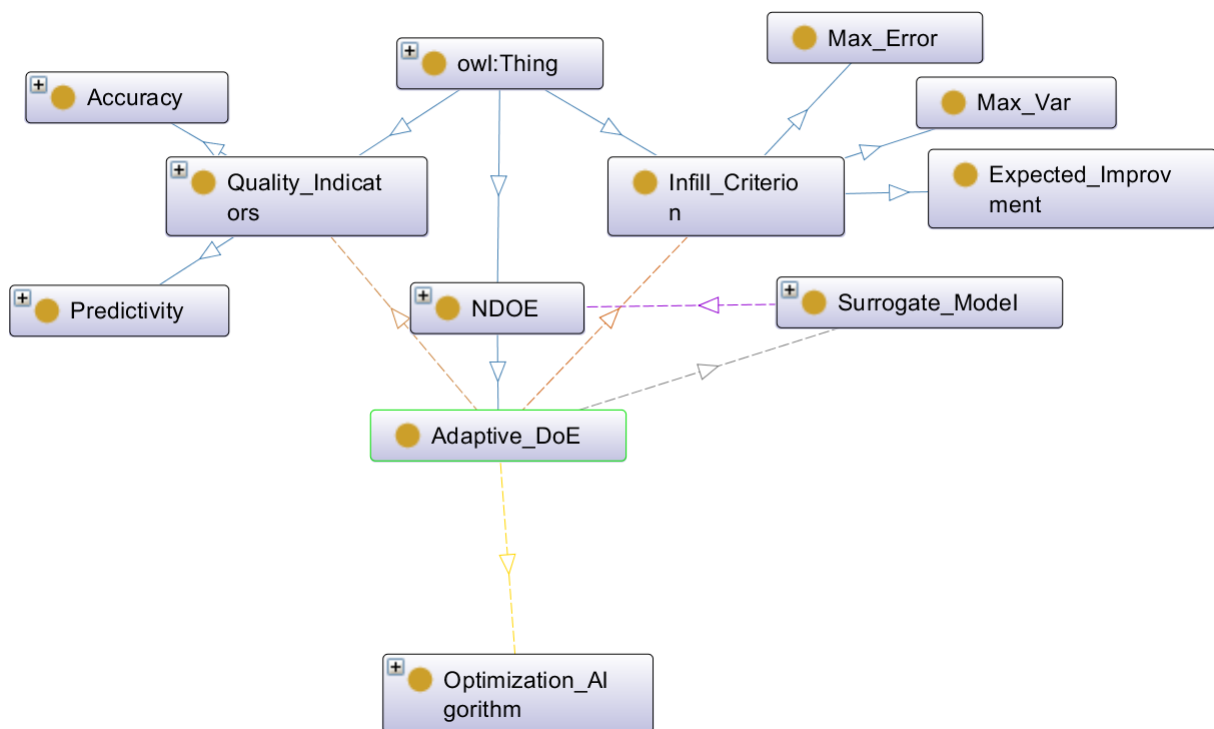
FIGURE 3.5 – Extrait de ODE : concepts de PEN et de processus de PEN



Les PEN adaptatifs (Figure 3.6) constituent un concept à part entière au sein de cette ontologie. La classe correspondante est une sous-classe de la classe (classe "NDOE") et hérite de ses propriétés. La définition d'un PEN adaptatif est complétée par les éléments spécifiques suivants :

- un unique algorithme d'optimisation, permettant d'explorer l'espace de conception à la recherche de l'expérience apportant le plus d'informations ;
- un unique critère de remplissage, permettant de quantifier l'apport d'une expérience ;
- plusieurs contraintes. Il peut s'agir, par exemple, d'un niveau de précision minimal pour un modèle de substitution ;
- un nombre d'expériences maximal à rajouter au PEN initial. Cette contrainte est ici considérée comme obligatoire, afin de se prémunir d'une explosion du coût de calcul en cas de non-convergences répétées de l'algorithme d'optimisation.

FIGURE 3.6 – Extrait de ODE : concepts de PEN adaptatif



## CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

Les propriétés liant les différentes classes de l'ontologie établissent des liens sémantiques. Par exemple, la proposition "un modèle de substitution est entraîné à partir d'un PEN" est traduit en OWL-DL<sup>2</sup> par le code présenté dans l'extrait 3.1.

Extrait de codes 3.1 – Exemple de propriétés en langage OWL (1/2)

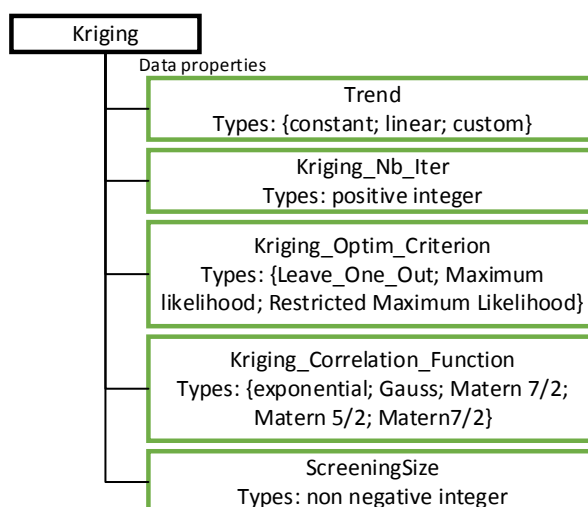
```
#http://www.semanticweb.org/gblondet/ontologies/2016/3/ODE#IsTrainedBy
:IsTrainedBy a owl:ObjectProperty ;
rdfs:domain :Surrogate_Model ;
rdfs:range :NDOE .
```

De même, des propriétés de données permettent de lier une classe à une donnée, tel qu'un nom ou un nombre. Par exemple, la classe "Kriging", sous-classe de "Surrogate\_ModelType" dispose de 5 propriétés de données, chacune définissant un paramètre à définir pour pouvoir utiliser un modèle empirique de type krigeage (Figure 3.7). L'assertion "un PEN est défini par un nombre d'expériences supérieur à 0" se traduit par le code présenté dans l'extrait 3.2.

Extrait de codes 3.2 – Exemple de propriétés de données en langage OWL (2/2)

```
#http://www.semanticweb.org/gblondet/ontologies/2016/3/ODE
#NumberOfExperiments
:NumberOfExperiments a owl:DatatypeProperty ;
rdfs:domain :NDOE ;
rdfs:range xsd:positiveInteger ;
rdfs:comment "Each NDoE embeds n experiments, n>0" .
```

FIGURE 3.7 – Détails des données de la classe Kriging.



2. avec la syntaxe Turtle

### 3.2.3 Retour sur les critères de conception de l'ontologie

La conception de cette ontologie repose sur les 5 critères de qualité d'une ontologie, formulés par [Gruber 93b] :

- une description sémantique claire, assurant une transmission du sens sans erreur ;
- une description sémantique cohérente, sans contradiction logique ;
- une capacité d'enrichissement par de nouveaux axiomes ou propriétés ;
- un biais d'encodage minimal : le fait de formaliser un domaine sémantique ne doit pas modifier la sémantique. En d'autres termes, le langage exprimant le domaine sémantique ne doit pas perturber le sens véhiculé ;
- un nombre d'axiomes minimal, afin que l'ontologie soit la plus générale et la plus réutilisable possible.

Il est nécessaire de trouver un compromis entre ces différents critères. En effet, la généralité d'une ontologie améliore sa capacité de ré-utilisation, mais diminue son utilité. De plus, une ontologie trop enrichie tend à diminuer sa clarté et rend sa cohérence plus difficilement maîtrisable. L'application de ces différents critères est décrite dans les paragraphes suivants.

#### Cohérence et clarté

L'utilisation du langage [OWL-DL](#) hiérarchise les concepts par classes et sous-classes, et définit les propriétés liant sémantiquement les classes entre elles. Les propriétés logiques intégrées dans l'ontologie permettent de contrôler les différentes instances et d'assurer la cohérence de la base de connaissances. Par exemple, une instance de la classe "NDoE" ne peut pas avoir, comme valeur de la donnée "NumberOfExperiments", une valeur négative ou nulle.

La cohérence logique de l'ontologie est contrôlée par des algorithmes spécifiques. L'algorithme Pellet [Sirin 07] a été utilisé au cours de la conception de cette ontologie. Il vérifie la syntaxe, la cohérence entre les classes et les propriétés, la cohérence entre l'ontologie et ses instances. Il permet également de déduire des relations non-spécifiées par inférences logiques à partir des propriétés existantes (Figure 3.8). L'algorithme Pellet supporte toutes les fonctionnalités du langage [OWL-DL](#), telles que la nature des propriétés (transitivité, propriété fonctionnelle, inverse, etc.) et les restrictions de cardinalité.

#### Extensibilité de l'ontologie

ODE a été définie de façon à disposer d'une richesse sémantique suffisante pour traiter les données et les connaissances utilisées dans le cadre du projet SDM4DOE. ODE comporte



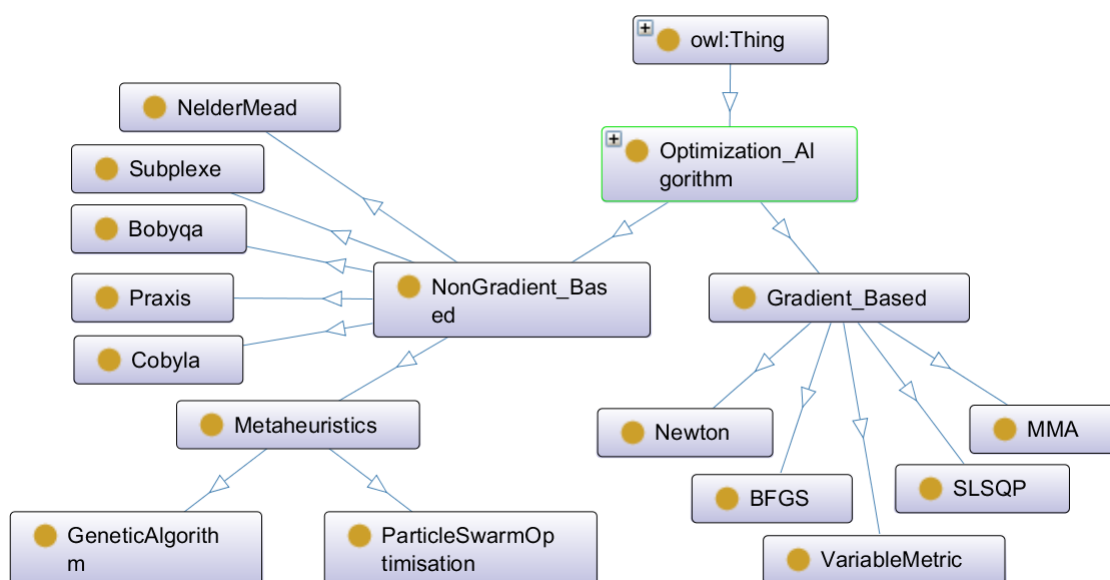
FIGURE 3.8 – Vérification de l'ontologie : résultats de l'algorithme Pellet

```

Log
INFO 09:49:41 REASONER CHANGED
INFO 09:49:43 ----- Running Reasoner -----
INFO 09:49:43 Pre-computing inferences:
INFO 09:49:43   - class hierarchy
INFO 09:49:43   - object property hierarchy
INFO 09:49:43   - data property hierarchy
INFO 09:49:43   - class assertions
INFO 09:49:43   - object property assertions
INFO 09:49:43   - data property assertions
INFO 09:49:43   - same individuals
INFO 09:49:44 Ontologies processed in 876 ms by Pellet
    
```

donc des concepts généraux, et des concepts spécialisés pour une application donnée. Il est nécessaire de distinguer les concepts clés (ex : les classes "NDoE" et "NDoE\_Process") des concepts spécialisés, liés à des domaines particuliers et des outils logiciels spécifiques. Ainsi, la classe "Optimization\_Algorithm" (Figure 3.9) dispose de trois sous-classes représentant les types principaux d'algorithmes d'optimisation ("Gradient\_Based", "Non-Gradient\_Based" et "Metaheuristics"), mais les sous-classes de ces types principaux sont spécifiques aux outils utilisés dans le cadre de ces travaux. Ces dernières ne constituent donc pas le cœur de l'ontologie transposable à d'autres domaines.

FIGURE 3.9 – Extrait de ODE : concepts d'algorithme d'optimisation



L'ontologie proposée ne redéfinit pas les concepts déjà définis dans d'autres ontologies. L'état de l'art (cf. chapitre 2) a permis d'identifier plusieurs ontologies et modèles de données décrivant des concepts liés au processus de PEN.

L'AP 209 de la norme ISO DIS 10303, intitulé "Multidisciplinary analysis and design", détaille, entre autres, les concepts liés à la simulation numérique. Ce protocole donne

une définition détaillée des modèles numériques utilisés dans le domaine de la mécanique, pour des études de mécanique des solides et des fluides. Il décrit également le lien entre ces modèles de simulation et les différents modèles géométriques. Il fournit la description des maillages, des résultats et des propriétés des produits. A titre d'exemple, un modèle par éléments-finis est décrit par un numéro d'identification, un nom, une description, les logiciels utilisés pour créer le modèle et exécuter la simulation, la description détaillée du maillage spatio-temporel associé, ainsi que la définition de chaque composant du produit et de leurs comportements respectifs. Chaque exécution est capitalisée et décrite par un identifiant, un nom, une description et les résultats de l'exécution. Cette description normalisée et détaillée du concept de modèle numérique initial, dans le cadre d'applications en mécanique, est ajoutée à l'ontologie proposée sous la forme d'une nouvelle classe, nommée "ISO\_10303\_AP209\_Analysis" liée à la classe "Finite\_element\_models", elle-même sous-classe de "Initial\_Models". D'autres concepts peuvent être enrichis par cette norme, tels que les données relatives aux projets, aux liens entre le processus de simulation, le processus de [PEN](#) et le processus de conception de produit.

PARO est une ontologie basée sur un modèle de données générique décrivant les relations entre les produits, les processus, les organisations et les ressources [Le Duigou 11]. La classe "Resources" intègre les ressources humaines, techniques, et méthodologiques. Cette description étend celle proposée initialement, et peut directement être liée à la classe "Resources" de l'ontologie proposée. D'autres concepts, tels que ceux décrivant les projets, enrichissent les concepts détaillés dans l'[AP 209](#).

L'ontologie EXPO [Soldatova 06] couvre le domaine des expériences scientifiques. EXPO décrit les concepts d'expériences physiques et numériques, de résultats expérimentaux, des tests statistiques permettant d'analyser les résultats et enfin les concepts liées à la traçabilité des expériences (auteurs, titre, etc.). Certains des concepts d'EXPO sont redondants avec ceux de l'ontologie proposée, tels que les concepts de facteurs (classe "Factor") et de variable d'intérêt (classe "TargetVariable"). La gestion des redondances est gérée par la propriété owl:equivalentClass spécifiée dans [OWL](#). EXPO est elle-même basée sur l'ontologie SUMO, décrivant des concepts plus généraux, tels que les unités de mesure, le temps, etc.

L'ontologie "Exposé" [Vanschoren 10], basée sur EXPO, porte sur les expériences réalisées spécifiquement pour des applications d'analyse de données (data mining). Elle définit les concepts d'ensemble de données (pour l'apprentissage), de paramètre, de prédiction, de ressources informatiques (au sens matériel). Elle spécifie également les processus et les flux (workflow) associés. Elle propose une description détaillée de toutes les principales méthodes d'analyse de données pour évaluer la prédictivité des modèles, effectuer des classifications, etc. Les classes "NDOE" et "Surrogate\_Model\_Type", par exemple, peuvent être étendues par les classes "dataset" et "predictive algorithm". En effet, un [PEN](#) exécuté

est un ensemble de données (dataset). Parmi les classes contenues dans la classe "predictive algorithm", les sous-classes "Gaussian processes" ou encore "neural network algorithm" sont présentes et décrivent respectivement les modèles de type krigeage et réseaux de neurones.

STATO<sup>3</sup> est une ontologie dédiée aux outils statistiques utilisés pour des applications en biologie, génomique et neuro-sciences. Cette ontologie comprend 678 classes et 53 propriétés. Une grande partie de ces classes sont spécifiques aux études biologiques. Certaines classes sont plus générales. Les classes "generically dependent continuant" et "planned process" contiennent des sous-classes définissant, de manière très détaillée, les concepts liés aux tests statistiques (tests d'hypothèses, analyse de la variance, etc.), aux mesures d'erreurs, aux lois de probabilités, à certains modèles de régression, aux plans d'expériences dédiés aux expériences physiques (ex : plan de Box-Behnken, plan composites, etc.) ou encore à la nature des variables. La classe "data transformation", sous-classe de "planned process" et couvrant de nombreuses méthodes d'analyse statistique, peut constituer une classe de référence pour les classes de ODE, notamment pour les classes "Statistical\_Distribution" et "Sensitivity\_Methods".

Les liens avec ces cinq ontologies ont été intégrés dans ODE (Figure 3.10). La Figure 3.11 illustre l'un des liens existant avec STATO entre la classe "NDoE\_Type" et la classe "study\_design" de STATO, marquant l'équivalence de ces deux concepts. Par cette équivalence, la classe "NDoE\_Type" couvre ainsi plus de types de PEN, déjà définis dans d'autres travaux, et complète la classe de STATO par de nouveaux types de PEN si besoin.

---

3. <http://bioportal.bioontology.org/ontologies/STATO>

FIGURE 3.10 – Représentation des liens inclus dans ODE

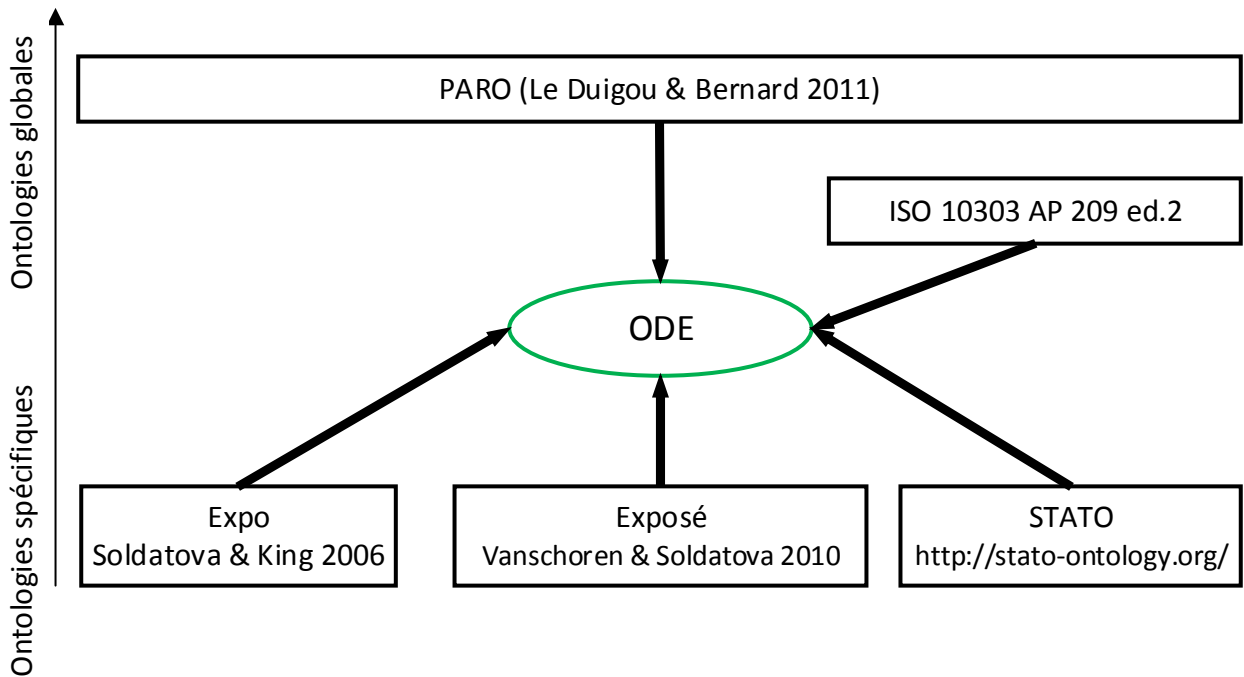


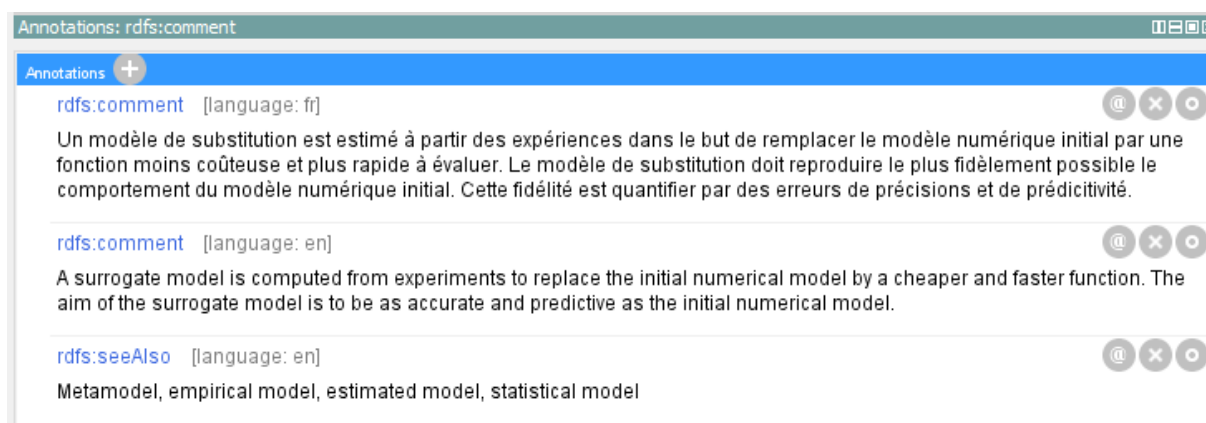
FIGURE 3.11 – Détails de la classe "NDoE\_Type" et lien avec l'ontologie STATO

The screenshot displays a software interface for ontology management. On the left, a 'Class hierarchy' tree shows 'owl:Thing' at the root, with 'NDoE\_Type' highlighted. Below it, several subclasses are listed, including 'Factorial', 'Space\_Filling\_Designs', and 'NDOE\_TYPE\_1'. On the right, the 'Usage: NDoE\_Type' pane shows 'Found 408 uses of NDoE\_Type'. It lists several uses, such as 'Factorial SubClassOf NDoE\_Type', 'hasNDoEType Range NDoE\_Type', and 'IsTypeOf Domain NDoE\_Type'. A 'Description: NDoE\_Type' section shows 'Equivalent To STATO\_study\_design'.

## Biais d'encodage et généricité

Les concepts décrit par ODE ont été définis selon l'état de l'art du domaine (cf. chapitre 1). Cette description entraîne une perte de généricité par le fait même d'exprimer ces connaissances par le langage. Pour minimiser l'impact de ce phénomène, des descriptions, commentaires et listes de synonymes et antonymes ont été ajoutés. Par exemple, le concept de modèle de substitution (classe "Surrogate\_Model") peut avoir plusieurs significations suivant la définition du concept de modèle. La définition est alors renseignée en commentaire, et les synonymes (métamodèle, modèle empirique, modèle statistique) sont associés. Cette solution peut être étendue en intégrant ces informations en plusieurs langues différentes (Figure 3.12).

FIGURE 3.12 – Extrait de ODE : annotations



### 3.2.4 Utilisation de l'ontologie

La structure de l'ontologie étant définie, son utilisation passe par la création d'instances des classes de ODE. Les Figures 3.13 et 3.14 donnent un aperçu de cette instanciation. Un processus de PEN, capitalisé dans la base de connaissances, est structuré par l'ontologie suivant chacune des classes pré-définies. Ici, le processus de PEN n°1 est associé à ses ressources, à un auteur particulier, dispose d'un objectif d'analyse de sensibilité, et se base sur un modèle de simulation par éléments finis. Le PEN associé comporte 43 expériences, 3 facteurs et une variable d'intérêt. Les autres caractéristiques de ce processus de PEN sont décrites par les autres classes de ODE.

FIGURE 3.13 – Exemple d'instance de la classe "NDOE\_Process"

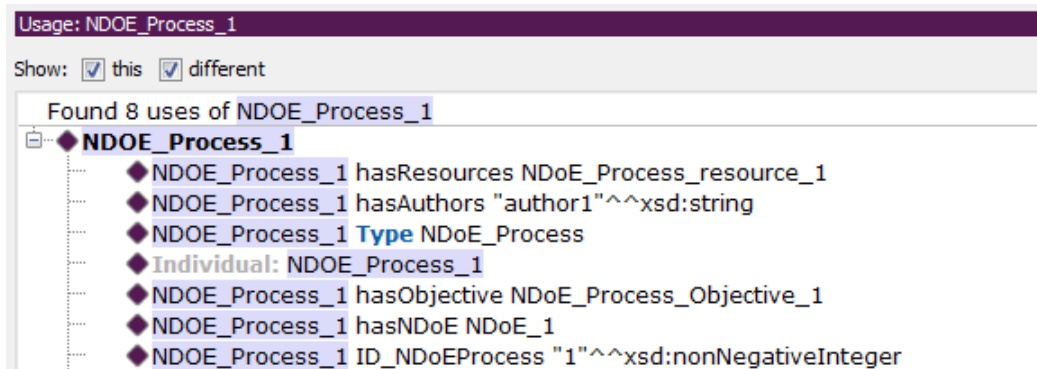
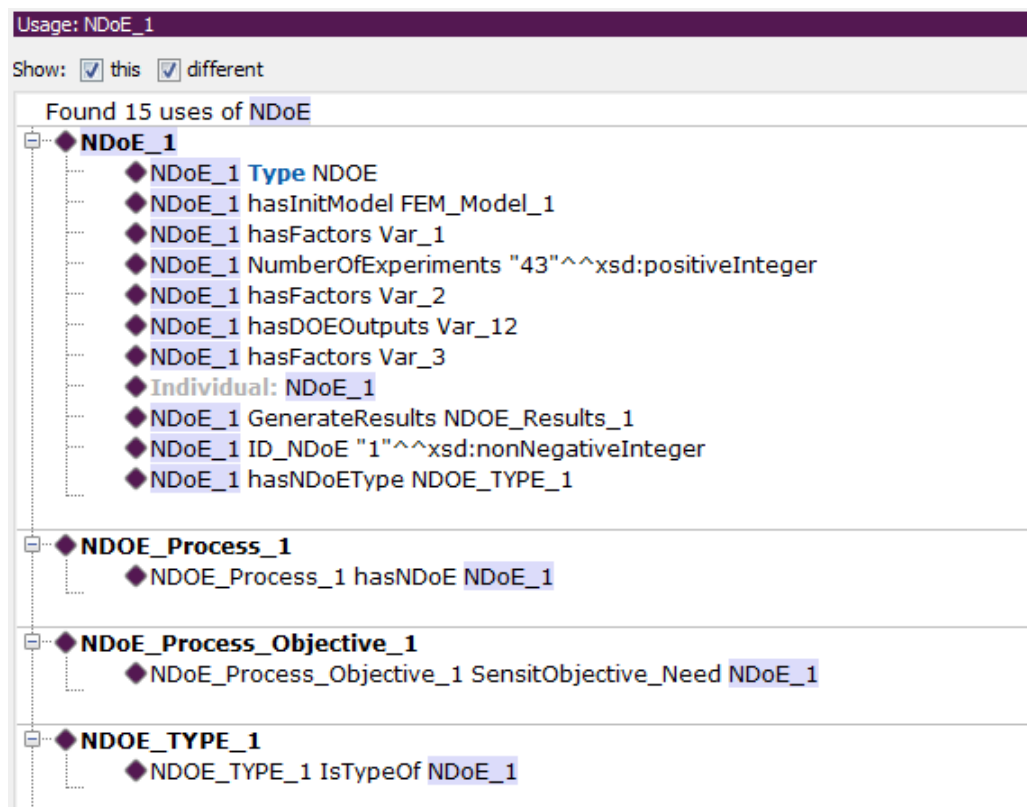


FIGURE 3.14 – Exemple d'instance de la classe "NDOE" liée à l'instance de la classe "NDOE\_Process"



### CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

Les instances sont définies séparément de la structure de ODE, afin de pouvoir réutiliser la structure indépendamment du cadre d'application. Une fois instanciée, ODE reçoit des requêtes en langage SPARQL pour extraire les données et les connaissances capitalisées. Par exemple, la requête présentée (Extrait de codes 3.3) extrait tous les processus de PEN stockés, ainsi que le type de PEN et le nombre d'expériences associées (Figure 3.15).

Extrait de codes 3.3 – Exemple de requête SPARQL

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ode: <http://www.semanticweb.org/gblondet/ontologies/2016/3/ODE#>

SELECT ?NDOE_Process ?Nb_Exp ?DoE_Type
WHERE {
?NDOE_Process ode:hasNDoE ?NDOE.
?NDOE ode:hasNDoEType ?NType.
?NDOE ode:NumberOfExperiments ?Nb_Exp.
?NType rdf:type ?DoE_Type
FILTER (?DoE_Type!= owl:NamedIndividual)
}
```

FIGURE 3.15 – Résultat de la requête SPARQL

The screenshot shows a web-based SPARQL query interface. At the top, there are several tabs: 'Active Ontology', 'Entities', 'Object Properties', 'Data Properties', 'Individuals by class', 'OWLviz', 'OntoGraf', 'Debugger', 'SPARQL Query', and 'custo'. The 'SPARQL Query' tab is active, displaying a query in a text area. Below the query, the results are shown in a table with three columns: 'NDOE\_Process', 'Nb\_Exp', and 'DoE\_Type'. The table contains 20 rows of data. At the bottom right of the interface, there is an 'Execute' button.

```

SPARQL query:
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ode: <http://www.semanticweb.org/gblondet/ontologies/2016/3/ODE#>

SELECT ?NDOE_Process ?Nb_Exp ?DoE_Type
WHERE {
  ?NDOE_Process ode:hasNDoE ?NDOE.
  ?NDOE ode:hasNDoEType ?NType.
  ?NDOE ode:NumberOfExperiments ?Nb_Exp.
  ?NType rdf:type ?DoE_Type
  FILTER (?DoE_Type != owl:NamedIndividual)
}
    
```

NDOE_Process	Nb_Exp	DoE_Type
NDOE_Process_362	"15"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_362	"15"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_69	"45"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Halton
NDOE_Process_69	"45"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_336	"38"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_336	"38"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Halton
NDOE_Process_399	"39"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_399	"39"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_32	"16"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_32	"16"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_160	"25"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_160	"25"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_108	"49"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_108	"49"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Halton
NDOE_Process_134	"12"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_134	"12"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type
NDOE_Process_197	"36"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	Sobol
NDOE_Process_197	"36"^^<http://www.w3.org/2001/XMLSchema#positiveInteger>	NDoE_Type



### **3.2.5 Apports de l'ontologie**

L'ontologie ODE a été proposée en réponse aux enjeux industriels et scientifiques d'application des méthodes de simulation numériques prenant en compte les sources de variabilités et d'incertitudes du produit. Les **PEN** apportent des éléments essentiels pour l'amélioration du produit au plus tôt dans le processus de développement du produit. En revanche, l'application d'un processus de **PEN** peut s'avérer être une tâche difficile pour des produits complexes. D'une part, une mauvaise configuration du processus de **PEN** conduit à un ralentissement des prises de décision. D'autre part, de nombreux choix sont nécessaires et nécessitent des connaissances expertes précises.

Cette proposition apporte :

- une structuration ontologique des connaissances spécifique aux **PEN**, jusqu'alors inexistante ;
- une intégration des connaissances relatives aux processus de développements de produits, aux statistiques, aux méthodes d'**IA** (Machine Learning, algorithmes d'optimisation) dans une même structure ontologique cohérente, par des liens entre les structures existantes et ODE ;
- une prise en compte des structures standards existantes spécifiées par l'**ISO** ;
- un support pour la capitalisation, le partage, la diffusion, la réutilisation et la création de connaissances relatives aux processus de **PEN**.

ODE participe à faciliter l'utilisation des processus de **PEN** dans l'industrie, en offrant un support pour les connaissances de l'entreprise en cohérence avec les ontologies existantes. Les connaissances sont alors mieux capitalisées, plus compréhensibles et plus facilement partagées dans un contexte d'ingénierie collaborative.

La section suivante décrit l'application d'un réseau bayésien, supporté par ODE. ODE structure les connaissances et les données de façon à les extraire et à former un ensemble d'apprentissage pour le réseau bayésien. Le réseau pourra alors assister l'utilisateur dans ses choix à partir des processus de **PEN** capitalisés dans l'entreprise étendue.

### **3.3 Réutilisation des connaissances capitalisées par réseaux bayésiens**

La transcription du domaine sémantique des processus de **PEN**, sous la forme d'une ontologie, vise à structurer les connaissances et les données pour les capitaliser et les partager. Cette structuration participe à accélérer les prises de décisions, notamment lors de la phase de configuration d'un processus de **PEN**. La seconde proposition présentée dans cette section traite de la réutilisation des connaissances pour assister l'utilisateur dans ses choix. Cela participe également à l'accélération des prises de décisions, tant durant le processus de **PEN** que durant les choix de conception réalisés à l'issue de l'analyse des résultats des **PEN**.

Compte-tenu du nombre de choix à faire, du nombre de possibilités pour chaque choix, de l'interdépendance de ces choix, ainsi que du fait que les connaissances soient incertaines et incomplètes, il est nécessaire de disposer d'une stratégie d'analyse des données capitalisées pour en extraire des connaissances. Cette analyse vise à identifier des configurations du processus de **PEN** efficaces (donnant les résultats désirés et suffisamment précis avec le minimum d'expériences) dans un contexte donné, et à terme, à confirmer ou infirmer les connaissances a priori, voir même à créer des nouvelles connaissances.

L'état de l'art a permis d'identifier le type de moteur d'inférences adapté pour remplir ces fonctions par rapport aux critères énoncés dans le chapitre 2, et notamment par rapport à leur capacité à mêler les connaissances expertes et l'analyse de données par apprentissage automatique.

Cette seconde partie décrit l'application des réseaux bayésiens pour la prédiction des configurations efficaces, ou pour prédire les performances d'une configuration donnée. La stratégie de conception du réseau bayésien est décrite, et les problématiques d'apprentissage de la structure et des paramètres du réseau sont abordés.

#### **3.3.1 Conception du réseau bayésien**

Une caractéristique importante, mise en évidence dans l'état de l'art (cf. chapitre 2), est sa capacité à intégrer les connaissances et les données dans un même modèle, sans être complètement dépendant de l'existence de connaissances a priori. Les prédictions du réseau bayésien résultent de la combinaison entre l'estimation de probabilités à partir des données et le calcul de probabilités à partir de lois pré-définies, provenant de l'expression des connaissances expertes. Cela permet d'enrichir des connaissances lacunaires par l'analyse des données, et dans le même temps, corriger les estimations à partir de données par les connaissances existantes.

Avant son utilisation, le réseau bayésien nécessite la définition de trois éléments :

1. les variables. Il faut choisir les variables de configuration à prendre en compte. Dans le cas présent, chaque variable est un élément de configuration du processus de **PEN**. Chaque variable représente un choix que devra faire l'utilisateur, selon le cas d'utilisation, avant d'exécuter le processus. Par exemple, il s'agit du type de **PEN**, du nombre d'expériences, du type d'objectif, du temps d'exécution du processus, etc. Toutes ne seront pas pertinentes suivant le problème traité. Par exemple, si l'objectif du processus de **PEN** est la production d'un modèle de substitution, il est inutile d'inclure la variable décrivant le choix d'une méthode d'analyse de sensibilité. Il s'agit de la première forme d'introduction de connaissances ;
2. la structure du réseau. Cette structure est le graphe acyclique dirigé représentant les liens de dépendance entre chaque variable. Cette structure peut être déterminée par l'apprentissage à partir des données, ou en interdisant/imposant un lien de dépendance entre deux variables. Il s'agit de la seconde forme d'introduction de connaissances au sein du réseau ;
3. les paramètres du réseau. Une fois la structure de dépendance établie, les probabilités conditionnelles sont déterminées. Là encore, elles peuvent être estimées à partir des données, ou imposées par l'utilisateur. Par exemple, la probabilité de choisir le degré d'un modèle empirique polynômial sachant que l'on a choisi un modèle de krigeage est nécessairement nul. Cette impossibilité logique se retrouvera dans les données d'apprentissage, mais peut être, dès le départ, imposée par une probabilité conditionnelle nulle. C'est la troisième forme d'introduction de connaissances.

Les sous-sections suivantes décrivent la détermination de ces trois éléments.

### **Le choix des variables**

La première étape de conception d'un réseau bayésien est la définition de ses variables. Chaque variable représente un choix à faire pour configurer le processus de **PEN**. Les variables à prendre en compte comprennent toutes les sous-méthodes et paramètres de ces méthodes impliquées dans la configuration du processus de **PEN** et qui influencent ses performances.

Une variable est définie par sa nature (qualitative ou quantitative, discrète ou continue) et par ses modalités. Par exemple, la variable "nombre d'expériences" est une variable discrète quantitative (nombre entier strictement positif). Elle peut, en théorie, disposer d'autant de modalités que de nombres entiers positifs existants, soit une infinité. Il faut donc définir une discrétisation résultant de deux compromis :

- le compromis entre le nombre de données à disposition et le nombre de modalités. Par exemple, on peut définir 3 modalités équilibrées : tous les processus comprenant entre 5 et 20 expériences, entre 21 et 50 expériences et ceux ayant plus de 50 expériences, contenant respectivement 30, 35 et 35 processus de PEN sur une base de 100 processus capitalisés. L'équilibrage favorise un apprentissage efficace. Une trop grande disparité pourrait faire chuter les performances prédictives du réseau pour certaines modalités. Si la première modalité ne dispose que d'un cas sur 100, le réseau bayésien aura plus de difficultés à faire de prédictions fiables que sur les 2 autres modalités disposant des 99 autres cas. A l'extrême inverse, ne considérer qu'une seule modalité englobant tous les nombres d'expériences existants ne permet pas de choisir et la variable devient inutile ;
- le compromis entre l'introduction de connaissances limitantes et la volonté d'explorer de nouvelles possibilités. Ce choix résulte de la connaissance du contenu de la base de données et des connaissances des utilisateurs. On ne cherche pas des plans de moins de 5 expériences si on sait qu'ils vont systématiquement mener à de mauvais résultats. A l'inverse, on peut volontairement laisser des modalités larges (ex : plus de 50 expériences) pour favoriser l'exploration de nouvelles solutions de configuration.

Il faut donc choisir les variables et les modalités selon l'intérêt qu'elles ont pour les utilisateurs et selon les connaissances et les données disponibles.

Les variables discrètes qualitatives couvrent les variables ayant un nombre fini de modalités non numériques. Par exemple, le choix du type de plans. Ce choix est limité aux types de plan existant et la discrétisation est immédiate (ex : modalité 1 : plan de Halton, modalité 2 : plan factoriel complet, etc.). Les deux compromis détaillés plus haut s'appliquent également. On peut écarter volontairement les modalités n'ayant pas été testées un nombre suffisant de fois par rapport aux autres, par précaution. On peut aussi limiter le nombre de modalités à celles qui paraissent intéressantes pour les utilisateurs, ou on peut toutes les considérer pour explorer toutes les possibilités.

Les variables continues sont supportées par les réseaux bayésiens. Elles sont représentées par des lois de probabilités continues, imposées par l'utilisateur ou estimées par apprentissage. La plupart des variables du processus de PEN sont discrètes. Seules quelques unes, tels que les seuils de convergence des algorithmes d'optimisation, sont réelles positives, et donc continues. Dans la pratique, il est courant de définir un seuil arrondi (ex :  $1 * 10^{-05}$ ), et donc, de discrétiser cette variable. En revanche, de nombreux résultats du processus de PEN sont des variables continues, telles que les indicateurs de précision et de prédictivité d'un modèle empirique ou le temps d'exécution du processus. Afin d'optimiser la définition de ce type de valeur, il peut être intéressant de considérer cette variable comme continue.

Enfin, le nombre de variables impacte directement la complexité du réseau, le coût de

calcul pour déterminer la structure et les paramètres, et les performances prédictives du réseau suite à l'apprentissage statistique. En trop grand nombre, toutes ces modalités peuvent rendre le réseau très complexe et l'apprentissage nécessitera de très nombreuses données et sera plus coûteux. Si il n'y en a pas assez, la capacité prédictive du réseau bayésien sera très limitée. Une méthode pour limiter le nombre de variable dans le réseau est de considérer une approche "multi-net". En effet, il est plus efficace d'avoir des réseaux peu complexes et performants qu'un seul réseau exhaustif [Naïm 07]. Cette approche profite du fait que certaines variables aient un caractère booléen pour partitionner le réseau. Par exemple, dans notre contexte, on peut former un réseau bayésien pour l'objectif de production d'un modèle de substitution, et un autre pour l'objectif d'analyse de sensibilité, sachant qu'une méthode d'analyse de sensibilité ne sera pas choisie pour un objectif de production d'un modèle de substitution, mais qu'elle le sera nécessairement pour une analyse de sensibilité. La variable "type d'objectif" est alors externalisée, et elle n'est plus représentée dans aucun réseau bayésien. Chaque réseau représentera une modalité de la variable "type d'objectif".

Les prédictions du réseau bayésien sont issues d'une combinaison entre une estimation statistique et une estimation probabiliste donnée a priori (prior bayésien). Ainsi, une modalité rare ne sera pas impossible. En cas de manque de données, la connaissance pourra contre-balancer le manque de données d'apprentissage. Si l'utilisateur n'a pas de connaissances précises et que les données sont en trop petit nombre, une loi de probabilité uniforme permet d'éviter de considérer certains choix comme impossibles. Si par exemple, aucun processus de [PEN](#) comportant 10 expériences n'a été testé, le réseau ne prédira pas que l'utilisation de 10 expériences est impossible. Sans connaissances ni données particulières, les modalités n'ayant jamais été testée seront équi-probables.

### **Structure du réseau bayésien**

La définition de la structure du réseau débute par la définition des dépendances et indépendances connues a priori. Cela se traduit par la définition d'arcs interdits et d'arcs imposés. Par exemple, on sait que le temps d'exécution est dépendant du nombre d'expériences réalisées. De même, on peut admettre que le niveau de prédictivité d'un modèle de substitution est indépendant du temps d'exécution. Ces deux variables peuvent être liées par le biais d'autres variables, mais il n'y a pas de lien direct. On peut donc interdire ce lien. Cette étape permet de simplifier la phase d'apprentissage de la structure.

Le reste de la structure est ensuite déterminé par apprentissage automatique sur la base des données disponibles. Il existe de nombreux algorithmes d'apprentissage, pouvant être classés en deux groupes principaux [Naïm 07].

Le premier groupe est basé sur la recherche d'indépendances conditionnelles, par le

biais de tests d'indépendances. Les deux principaux algorithmes sont les algorithmes PC [Spirtes 93] et IC (Inductive Causation) [Pearl 00]. Le premier part d'une structure comportant tous les arcs possibles, pour en supprimer certains à la suite des tests d'indépendance, alors que l'algorithme IC part d'une structure vide et rajoute progressivement les arcs. Une recherche brute de la meilleure structure parmi l'ensemble des structures possibles à un coût super-exponentiel selon le nombre de variables. Par exemple, pour 10 variables, il a été démontré [Robinson 77] qu'il existe  $4,2 * 10^{18}$  structures possibles. Il n'est donc pas envisageable de mener une telle recherche. L'algorithme PC limite le coût en effectuant les tests d'indépendances par ordre croissant d'ordre de dépendance (ordre 0 :  $A \perp B$  ; ordre 1 :  $A \perp B \mid C$ , etc.). De multiples variations de ces algorithmes existent afin de réduire le coût de l'apprentissage [Naïm 07, Denis 14].

Le second groupe rassemble les approches basées sur un score que l'on cherche à maximiser. Ce score quantifie l'adéquation aux données d'apprentissage. Les algorithmes appartenant à ce groupe diffèrent selon le score utilisé et la méthode de recherche de structures. Comme pour les algorithmes du premier groupe, il est impossible, en un temps raisonnable, d'évaluer le score de toutes les structures possibles. Plusieurs scores existent, on peut citer celui d'entropie conditionnelle, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) ou  $BD\gamma$  (Bayesian Dirichlet) [Naïm 07]. Certains de ces scores prennent en compte la complexité du modèle (par exemple, les scores AIC et BIC). Le score est maximisé si l'adéquation aux données et la simplicité de la structure sont maximisées. Cela amène toutefois à une certaine insensibilité par rapport à certaines dépendances. Les dépendances faibles augmentent plus fortement la complexité de la structure que l'adéquation aux données. Elles sont donc ignorées.

La définition de l'espace de recherche, contenant les structures évaluées ou les dépendances analysées, peut être définie par différentes techniques. Cette espace peut être restreint, en appliquant le principe d'équivalence de la structure au sens de Markov. Plusieurs structures apportent les mêmes résultats, même si elles ne reflètent pas la causalité réelle. D'autres heuristiques sont applicables pour restreindre cet espace de recherche, telles que l'ordonnement des nœuds de la structure couplés au calcul d'un score, ou encore la recherche gloutonne [Russel 10].

La Figure 3.16 illustre le résultat de l'apprentissage d'un réseau dans le cas du modèle numérique T1 présenté au chapitre 1. Cette figure illustre également l'influence de l'ensemble d'apprentissage sur la structure finale. Dans ce cas précis, aucune connaissance experte n'a été intégrée. L'apprentissage de la structure est assuré par l'algorithme Hill-Climbing ([Denis 14]), basé sur une recherche gloutonne des arcs de dépendances, et sur un score BIC. Cet algorithme présente l'avantage de donner automatiquement une structure dirigée (les arcs sont orientés). Beaucoup d'autres algorithmes nécessitent un post-traitement manuel pour diriger les arcs. Le réseau présenté tient compte des variables de configurations

de **PEN** adaptatifs pour obtenir un modèle de substitution du modèle T1. Les variables considérées ici couvrent (cf.chapitre 1) :

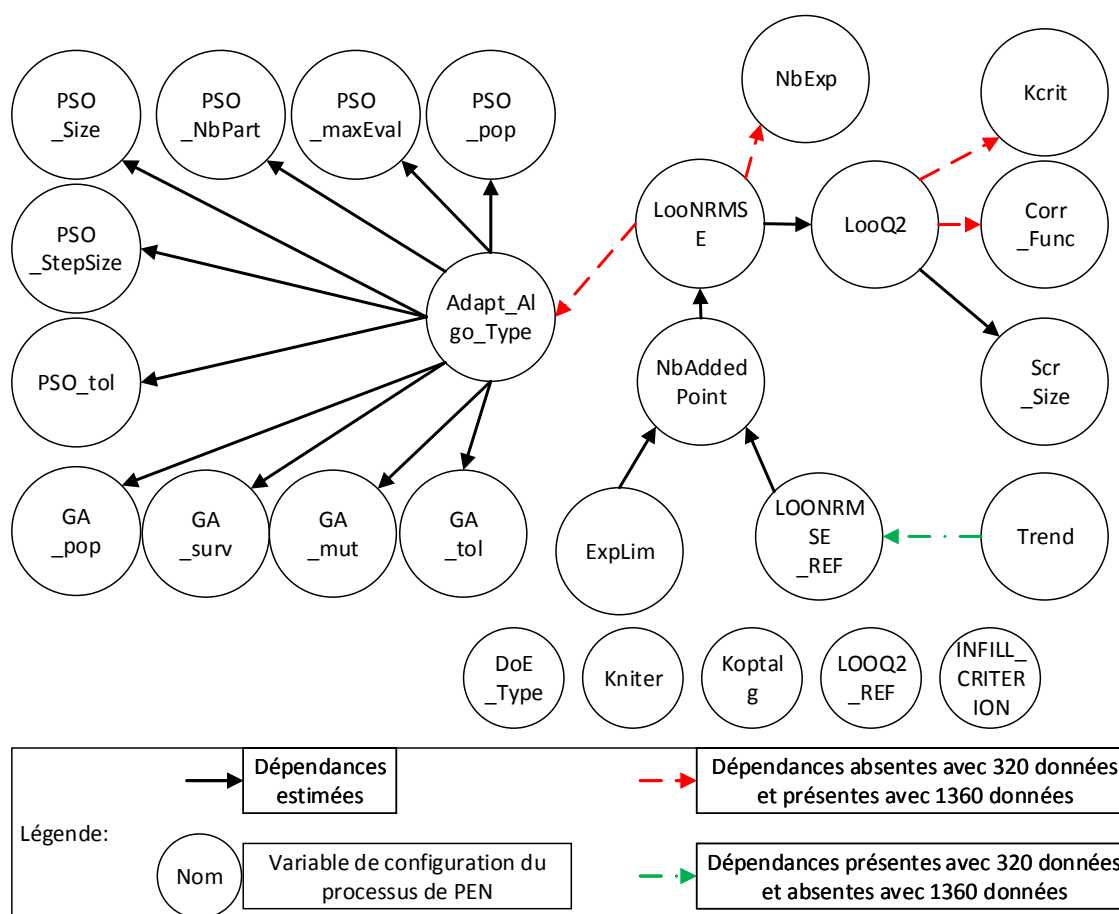
- le type de **PEN** ("DoE\_Type");
- le nombre d'expériences ("Nb\_Exp");
- les paramètres du modèle de substitution de type Krigeage ("Kcrit", "Corr\_Func", "Scr\_Size", "Trend", "Koptalg", "Kniter");
- le type de métaheuristique pour trouver les expériences enrichissant le **PEN** ("Adapt\_Algo\_Type"), les 6 paramètres de l'algorithme d'essaim particulaire "PSO" et les 4 paramètres de l'algorithme génétique "GA";
- le nombre d'expériences ajoutées au **PEN** initial ("NbAddedPoint");
- le nombre maximal d'expériences que l'on se permet d'ajouter au **PEN** (ExpLim");
- le critère de remplissage ("INFILL\_CRITERION") pour quantifier l'apport d'une nouvelle expérience au **PEN** initial;
- les objectifs de précision et de prédictivité du modèle de substitution de krigeage ("LOONRMSE\_REF" et "LOOQ2\_REF");
- la précision et la prédictivité effectives du modèle de substitution produit à l'issue du processus de **PEN** ("LOONRMSE" et "LOOQ2").

Le graphe obtenu permet à l'utilisateur de visualiser le réseau et de comprendre rapidement les variables dont le choix aura un effet significatif dans la performance du processus de **PEN**. Dans ce cas précis, on remarque que le type de **PEN** n'a pas d'influence, et pourra être choisi arbitrairement sans risque. La Figure 3.16 montre également la nécessité d'inclure des connaissances dans le réseau, si elles sont disponibles. Certaines dépendances (arcs rouges) n'apparaissent que lorsque qu'un très grand nombre de processus de **PEN** a été capitalisé (1360). A priori, on pouvait légitimement supposé un lien entre la précision finale du modèle de substitution de krigeage et le nombre d'expériences, par exemple.

La stratégie appliquée dans cette proposition est l'approche "multi-net". Au lieu d'un réseau bayésien unique couvrant l'intégralité des variables de configuration du processus de **PEN**, plusieurs réseaux, comportant moins de variables, sont définis. Des réseaux différents sont utilisés suivant le type de produit, le type modèle numérique et le phénomène physique simulé, le type de démarche de **PEN** (adaptatif ou non) et l'objectif (production d'un modèle de substitution, optimisation, analyse de sensibilité, etc.).

Une fois la structure obtenue, la détermination des paramètres du réseau, c'est à dire des probabilités conditionnelles, sont déterminées.

FIGURE 3.16 – Evolution de la structure du réseau bayésien pour le modèle T1 en PEN adaptatif. Apprentissage Hill-Climbing. Passage de 1360 données à 320 données d'apprentissage.





### Paramètres du réseau bayésien

Une fois la structure déterminée, la seconde phase de construction d'un réseau bayésien consiste à déterminer les probabilités conditionnelles associées. De la même façon que pour la détermination de la structure, les connaissances a priori sont d'abord intégrées. On peut imposer un type de loi de probabilité ou imposer certaines probabilités, selon les connaissances expertes disponibles.

La méthode d'apprentissage la plus simple pour l'apprentissage des paramètres est l'utilisation de l'estimateur du maximum de vraisemblance, qui approche la probabilité d'un évènement par la fréquence d'apparition de cet évènement. Cet estimateur est présentée par l'équation 3.1 reprise de [Naïm 07]. La probabilité que la variable  $X_i$  prenne la valeur  $x_j$  sachant que les variables parentes de  $X_i$  sont dans l'état  $x_j$  est approchée par un estimateur  $\theta$  égale au rapport entre le nombre d'occurrences de l'évènement  $X_i = x_k | p(X_i) = x_j$ , noté  $N_{i,j,k}$ , et l'ensemble des données de la base d'apprentissage  $E_{app}$ .

$$\hat{P}(X_i = x_k | p(X_i) = x_j) \simeq \hat{\theta}_{i,j,k}^{MV} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \quad (3.1)$$

L'apprentissage bayésien est, quant à lui, une autre approche permettant d'exprimer d'autres estimateurs des probabilités conditionnelles. En supposant que la distribution des données suit une loi multinomiale, deux estimateurs peuvent être définies : l'estimateur de l'espérance a posteriori (EAP : équation 3.3) et l'estimateur du maximum a posteriori (MAP : équation 3.4) [Naïm 07]. Ces équations sont basées sur le théorème de Bayes (équation 3.2), avec  $\alpha_{i,j,k}$  les paramètres de la distribution de Dirichlet (en supposant que la distribution de l'ensemble d'apprentissage suit une loi multinomiale) associée à la loi a priori de  $\hat{P}(X_i = x_k | p(X_i) = x_j)$ .

$$P(\theta | E_{app}) \propto P(E_{app} | \theta) P(\theta) \quad (3.2)$$

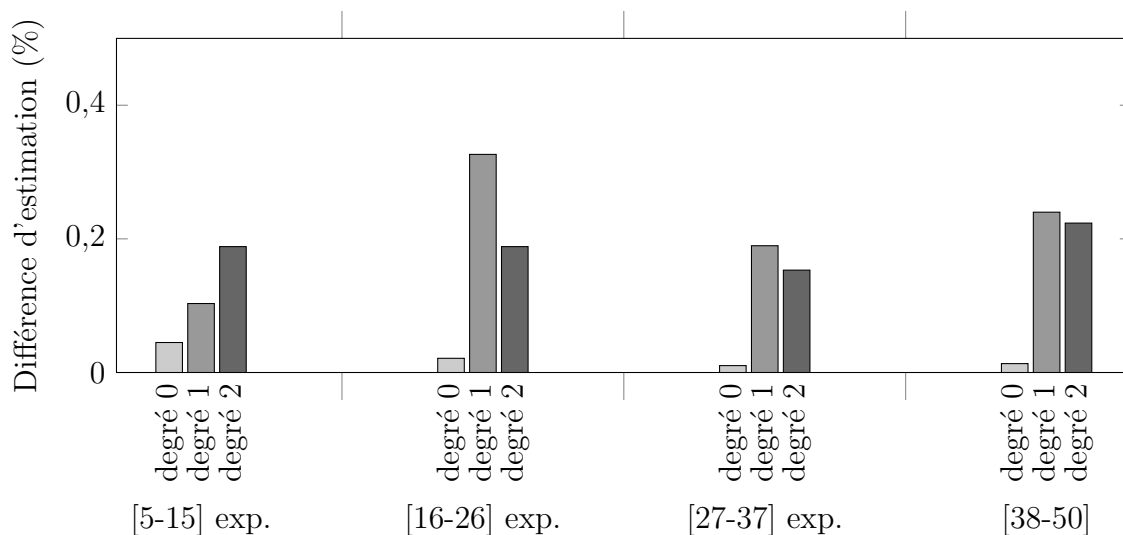
$$\hat{\theta}_{i,j,k}^{EAP} = \frac{N_{i,j,k} + \alpha_{i,j,k}}{\sum_k N_{i,j,k} + \alpha_{i,j,k}} \quad (3.3)$$

$$\hat{\theta}_{i,j,k}^{MAP} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k N_{i,j,k} + \alpha_{i,j,k} - 1} \quad (3.4)$$

Enfin, l'algorithme EM (Expectation Maximisation) peut être utilisé pour estimer les paramètres, notamment dans le cas où certaines données seraient manquantes [Fahad 14]. Dans le chapitre 4, traitant un cas d'utilisation concret, les paramètres sont estimés par une méthode d'apprentissage bayésien [Denis 14] spécifique, qui dans ce cas combine un

estimateur statistique et un prior bayésien (loi de probabilité issues de la connaissance experte). Il est possible dans ce cas de régler l'importance de l'un ou de l'autre suivant la quantité de données ou de connaissances à disposition. L'utilisation du maximum de vraisemblance et de la méthode d'apprentissage bayésien utilisé dans le chapitre 4 aboutit sensiblement aux mêmes résultats lorsque le réglage de la méthode bayésienne est fait de manière à privilégier les données par rapport aux connaissances (Figure 3.17). Cet histogramme présente la différence d'estimation des probabilités conditionnelles de choix d'un type de modèle de substitution (krigeage et chaos polynômial de degré 1 ou 2) suivant le nombre d'expériences défini au préalable. Cette différence entre les deux méthodes d'estimation ne dépasse pas 0.4%, ce qui montre que dans cette application, les données ont été privilégiées par rapport aux connaissances.

FIGURE 3.17 – Différence d'estimation des paramètres (probabilités) du réseau bayésien entre le maximum de vraisemblance et la méthode bayésienne (valeurs absolues).



Probabilités conditionnelles de choix du degré du modèle de chaos polynômial connaissant le nombre d'expériences du PEN. Le degré 0 indique le choix d'un autre type de modèle empirique

Le maximum de vraisemblance estime de nombreuses probabilités comme étant strictement nulles lorsqu'un évènement n'est jamais survenu. Ainsi, si une configuration n'a jamais été testée, elle ne pourra jamais être prédite. La méthode bayésienne n'estime que des probabilités strictement positives, ce qui rend les inférences moins catégoriques. Cet exemple a été obtenu avec le modèle numérique T1 (cf.chapitre 1) pour un objectif de production d'un modèle de substitution avec un **PEN** non-adaptatif.

Afin de vérifier l'apprentissage de la structure et des paramètres, des méthodes de validation croisées sont mises en œuvre, à l'instar des modèles de substitutions. L'objectif est alors de savoir si l'on peut faire confiance aux prédictions du réseau bayésien. Ces méthodes sont présentées dans la suite de ce chapitre.

### 3.3.2 Validation des performances prédictives du réseau bayésien

Une première validation croisée (cf. chapitre 1) par la méthode "k-fold" a été appliquée [Hastie 09]. Cette méthode consiste à partitionner l'ensemble d'apprentissage en plusieurs sous-ensembles de même taille. Un sous-ensemble est utilisé comme ensemble de validation, et les autres sont utilisés pour l'apprentissage. Dans cette application, cette opération est répétée  $n$  fois avec  $k$  sous-ensembles de validation différents. Pour chaque répétition, un nouvel ensemble d'apprentissage, et un nouvel ensemble de validation, sont formés, un apprentissage des paramètres est effectué et une erreur de prédiction est calculée sur l'ensemble de validation. Cette validation croisée porte uniquement sur l'estimation des paramètres (probabilités) du réseau bayésien. La structure est produite une fois pour toutes et elle est supposée insensible aux changements de l'ensemble d'apprentissage opérés lors de la validation croisée. Les différents ensembles d'apprentissage contiennent toujours 80% des données disponibles, mais sans aucune garantie qu'il puisse s'agir des mêmes données dans chaque ensemble.

L'erreur calculée à chaque itération est le rapport entre le nombre de mauvaises prédictions<sup>4</sup> et le nombre total de données de validation. L'erreur global est la moyenne des  $n$  erreurs calculées. Cette erreur est calculée pour chaque variable, à partir des 2000 données disponibles. La valeur de  $k$  est choisie de façon à avoir constamment 1600 données d'apprentissage (soit 80% des données) à chaque exécution, ce qui donne  $k=5$ . Le processus de validation est répété 100 fois ( $n=100$ ). La répartition des erreurs globales sur les 100 itérations est représentée en Figure 3.18 pour la variable LooQ2, représentant la prédictivité du modèle de substitution. L'erreur moyenne de classification est de 0.1133, avec un écart-type de 0.0005. La performance prédictive du réseau bayésien est donc peu sensible à l'ensemble d'apprentissage, et l'erreur est faible.

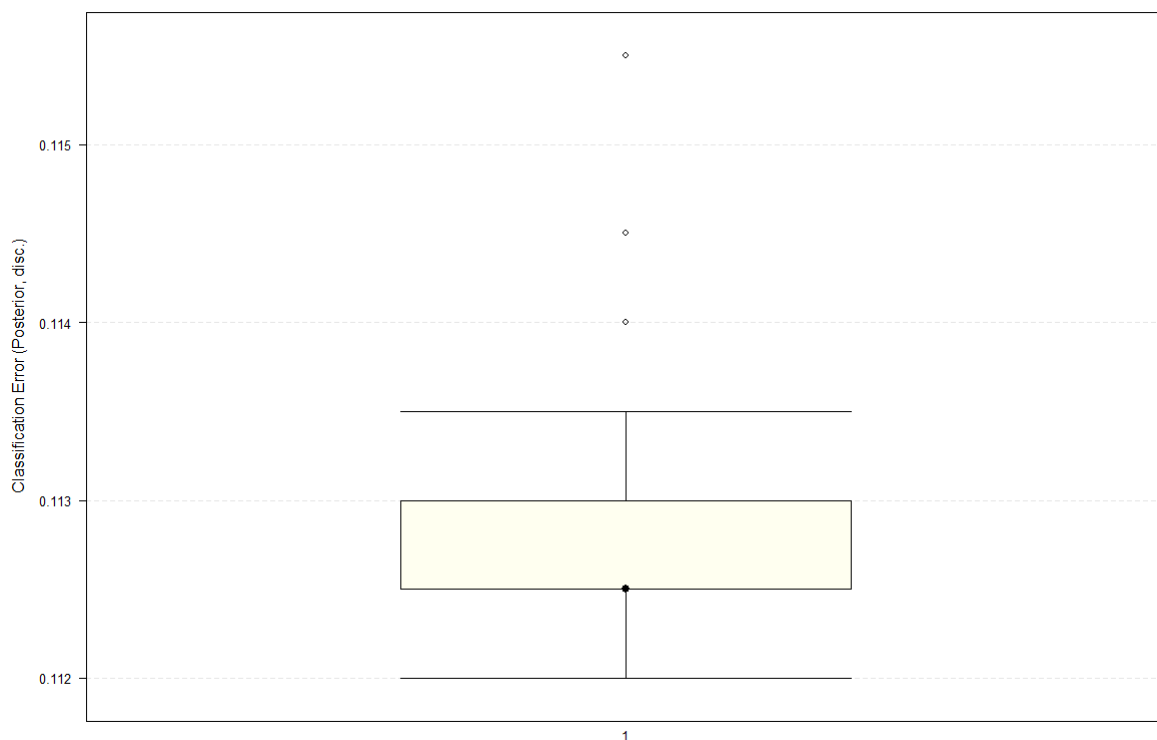
Une seconde validation croisée plus détaillée est ensuite mise en œuvre. Cette validation se base sur un unique ensemble d'apprentissage et un unique ensemble de validation, contrairement à la première validation croisée. Les ensembles d'apprentissage et de validation sont dans ce cas strictement identiques à ceux définis lors de l'apprentissage de la structure. Cette seconde démarche permet d'analyser de façon détaillée les capacités prédictives du réseau bayésien issu de l'apprentissage.

Pour chaque variable, une matrice de confusion est générée [Powers 11]. Elle représente les différentes prédictions réalisées par le réseau bayésien pour une variable donnée. Les éléments diagonaux de cette matrice représentent le nombre de configurations correctement classées. Les éléments extra-diagonaux contiennent le nombre des configurations mal

---

4. ou mauvaises classifications, survenant lorsque le réseau bayésien prédit à tort, par exemple, qu'une configuration de processus de PEN atteigne une certaine performance

FIGURE 3.18 – Erreurs de prédiction du réseau bayésien pour la variable  $LooQ^2$ , estimées par validation croisée de type "k-fold",  $k=5$



classées. Pour le cas présent, la matrice de confusion suivante (Tableau 3.1) est obtenue pour la variable  $NbAddedPoint$  à partir de la structure présentée par la Figure 3.16.

TABLEAU 3.1 – Matrice de confusion du réseau bayésien, obtenus pour le modèle T1, pour la variable  $NbAddedPoint$  (nombre d'expériences ajoutées au cours du PEN adaptatif) après un apprentissage sur 320 données à partir de la structure précédemment obtenue (Figure 3.16).

NbAddedPoint	Modalités	Prédictions			
		[1-10]	[11-15]	0	FAIL
Observations	[1-10]	36	9	0	0
	[11-15]	0	21	0	0
	0	4	9	0	0
	FAIL	0	1	0	0

Cette matrice de confusion montre par exemple que, dans l'ensemble de validation, 45 processus de PEN ont eu entre 1 et 10 expériences ajoutées au cours du processus adaptatif (somme des éléments de la première ligne,  $36 + 9 + 0 + 0$ ), mais seulement 36 de ces observations ont été prédites correctement par le réseau bayésien basé sur l'ensemble d'apprentissage. Il y a donc eu 36 vrais positifs (observations et prédictions concordantes, noté VP), 9 faux négatifs (cas observés dans la modalité [1-10] mais prédit dans une autre modalité, noté FN), 4 faux positifs (prédictions d'appartenance à la modalité [1-10] mais observations dans une autre modalité, noté FP) et enfin 31 vrais négatifs (cas observés dans les toutes les autres modalités, et prédits également dans les autres modalités, noté

VN).

A partir de chaque matrice de confusion générée pour chaque variable du réseau bayésien, des indicateurs sont calculés pour évaluer ses performances prédictives. Ces indicateurs sont utilisés de façon habituelle afin de contrôler les modèles de classification, et il existe d'autres indicateurs non développés dans ce chapitre [Powers 11]. Les indicateurs calculés ici sont (voir les équations 3.5 à la page suivante) :

- la répartition des données d'apprentissage et de validation par modalité. Il est nécessaire de vérifier si ces ensembles sont équilibrés, globalement et localement pour chaque variable et chaque modalité. Si la répartition des données dans l'ensemble des modalités est différente entre l'ensemble d'apprentissage et l'ensemble de validation, le réseau bayésien pourrait ne pas être validé correctement et ne pas être capable de prédire correctement. De plus, si une modalité est très peu représentée dans l'ensemble d'apprentissage par rapport aux autres, les performances prédictives du réseau bayésien pour cette modalité risquent d'être mauvaises (surtout si l'apprentissage statistique est prépondérant par rapport au prior bayésien) ;
- le taux de vrais positifs (TPR, équation 3.5a), appelé aussi "Sensitivity" ou "Recall" dans le domaine du Machine Learning. Il donne la probabilité de prédire qu'une configuration de processus de PEN appartienne à une modalité et qu'il appartient réellement à cette modalité. Cet indicateur donne un niveau de confiance dans la prédiction ;
- le taux de vrais négatifs (TNR, équation 3.5b), ou "Specificity". Il donne la probabilité de prédire qu'une configuration de processus de PEN n'appartienne pas à une modalité et qu'elle n'appartient effectivement pas à cette modalité. De la même façon que pour le Recall, il donne une idée de la confiance que l'on peut prêter au réseau bayésien ;
- le taux de faux positifs (FPR, équation 3.5c). Cet indicateur se déduit directement du taux de vrais négatifs ;
- le taux de prédictions positives (PPV, équation 3.5d). Il correspond au rapport entre le nombre de vrais positifs prédits et le nombre de prédictions positives ;
- le taux de prédictions négatives (TPN, équation 3.5e). Il correspond au rapport entre le nombre de vrais négatifs prédits et le nombre de prédictions négatives ;
- la mesure de précision globale, ou "Accuracy" (équation 3.5f) indique le rapport entre le nombre de bonnes prédictions et le nombre total de prédictions. C'est un indicateur global qui, s'il n'est pas associé aux autres indicateurs, peut s'avérer trompeur. Un modèle prédictif peut commettre une erreur globalement faible, mais être très mauvais sur une modalité en particulier ;
- l'erreur moyenne globale de classification, obtenue par la méthode de validation croisée "k-fold", utilisée précédemment. Associé à son écart-type, il donne une mesure

de la sensibilité de l'erreur par rapport à l'ensemble de validation utilisé.

$$TPR = \frac{VP}{VP + FN} \quad (3.5a)$$

$$TNR = \frac{VN}{VN + FP} \quad (3.5b)$$

$$FPR = 1 - TNR \quad (3.5c)$$

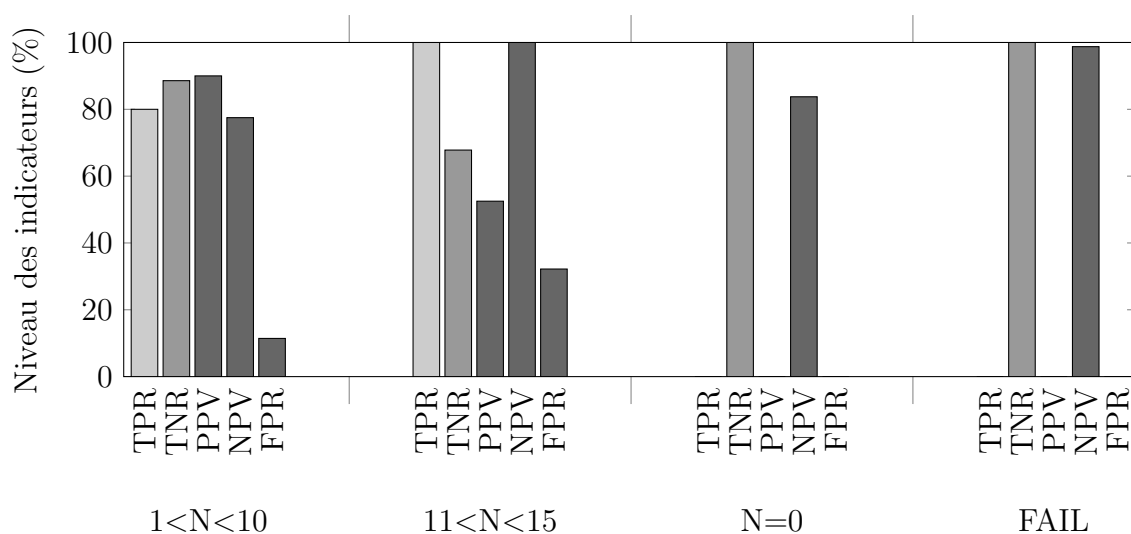
$$PPV = \frac{VP}{VP + FP} \quad (3.5d)$$

$$TPN = \frac{VN}{VN + FN} \quad (3.5e)$$

$$Accuracy = \frac{VN + VP}{VN + FN + VP + FP} \quad (3.5f)$$

Le calcul de ces indicateurs est développé dans le Tableau 3.2, à partir de l'exemple utilisé précédemment. Ces résultats sont également synthétisés par la Figure 3.19.

FIGURE 3.19 – Histogramme des indicateurs de prédictivités du réseau bayésien pour la variable NbAddedPoint.



Afin d'expliciter ces résultats, les interprétations suivantes sont données à titre d'exemple ;

- la répartition des données d'apprentissage et de validation par modalité. Ces deux ensembles ne sont pas équilibrés ici. Par exemple, les données de la modalité "[1-10]" représentent la majorité de l'ensemble d'apprentissage. On peut s'attendre alors que le réseau bayésien donne de moins bonnes prédictions pour les autres modalités. On remarque aussi que les ensembles d'apprentissage et de validation sont cohérents : la proportion de processus de PEN dans chaque modalité est sensiblement identique ;
- la probabilité qu'un processus de PEN soit prédit dans "[11-15]" et que c'est effectivement le cas est de 1 (TPR = 100%). Ainsi, pour 100 observations positives

### CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

TABLEAU 3.2 – Matrice de confusion et indicateurs de performance prédictive du réseau bayésien, obtenus pour le modèle T1, pour la variable NbAddedPoint (nombre d'expériences ajoutées au cours du PEN adaptatif) après un apprentissage sur 320 données à partir de la structure précédemment obtenue (Figure 3.16).

NbAddedPoint	Modalités	Prédictions			
		[1-10]	[11-15]	0	FAIL
Observations	[1-10]	36	9	0	0
	[11-15]	0	21	0	0
	0	4	9	0	0
	FAIL	0	1	0	0
<b>Ensembles de données utilisés</b>					
Nombre de données d'apprentissage		320			
Nombre de données de validation		80			
<b>Indicateurs par modalités</b>					
Répartition des données d'apprentissage (/320)		56.25%	26.88%	14.69%	2.19%
Répartition des données de validation (/80)		56.25%	26.25%	16.25%	1.25%
Nombre de vrais positifs (VP)		36	21	0	0
Nombre de faux positifs (FP)		4	19	0	0
Nombre de vrais négatifs (VN)		31	40	67	79
Nombre de faux négatifs (FN)		9	0	13	1
TPR ("Recall")		80.00%	100%	0%	0%
TNR ("Specificity")		88.57%	67.80%	100%	100%
FPR		11.43%	32.20%	0%	0%
PPV ("Precision")		90.00%	52.50%	N/A	N/A
NPV		77.50%	100%	83.75%	98.75%
<b>Indicateurs globaux</b>					
Accuracy		71.25%			
Erreur moyenne globale de classification (k-fold)		23.16% (écart-type :0.0034)			

(appartenant à cette modalité), ces mêmes 100 processus de PEN seront prédits correctement. On peut donc être presque certain que si le réseau bayésien donne une prédiction dans cette modalité, pour cette variable, il ne se trompe pas. Si l'utilisateur lance un processus de PEN en suivant les conseils du réseau bayésien, qui lui prédit que le processus n'ajoutera que entre 11 et 15 expériences numériques, il est quasiment certain que ce soit réellement le cas, et l'utilisateur pourra prévoir l'utilisation de ses ressources ;

- la probabilité qu'un processus de PEN ne soit pas dans la modalité "[11-15]" et que cela s'avère effectivement le cas est de 0.678 (TNR=67.8%). Cela implique un taux de faux positifs important (FPR=32.2%), conduisant à prédire un processus de PEN adaptatif comme appartenant, à tort, dans la modalité "[11-15]";
- pour 100 prédictions positives ("le processus de PEN appartient à cette modalité"), 52 de ces prédictions sont en effet positives (PPV=52.5%);
- la valeur de NPV est maximale (NPV = 100%), ce qui signifie que toutes les

prédictions négatives sont vraies. Ainsi, on est certain que le réseau bayésien ne commettra pas d'erreur s'il prédit qu'un processus de **PEN** n'appartient pas à la modalité "[11-15]";

- les indicateurs globaux montre un taux d'erreur non-négligeable, entre 20% et 30%. Cela donne un rapide aperçu de la qualité des prédictions.

Les indicateurs présentés ci-dessus sont ensuite calculés pour toutes les variables. Ainsi, à la suite de l'apprentissage, l'utilisateur dispose d'une cartographie complète des performances prédictives attendues du réseau bayésien, par rapport aux caractéristiques de l'ensemble de validation et de l'ensemble d'apprentissage.

A l'issue de la phase de vérification, l'utilisateur réalise ses choix finaux et exécute le processus de **PEN**. Une fois les résultats obtenus, toutes les données (configuration, résultats, et performances du processus de **PEN**) sont ajoutées à la base de connaissances par le biais de l'ontologie ODE.

### **3.3.3 Apports du réseau bayésien**

L'application d'un réseau bayésien pour la réutilisation des connaissances relatives aux processus de **PEN** apporte au système à base de connaissances plusieurs capacités :

- le système intègre les connaissances expertes existantes dans l'entreprise sous forme de loi de probabilité a priori ou de dépendances interdites ou obligatoires entre les différents éléments de configuration d'un processus de **PEN** (variables du réseau bayésien) ;
- le système est également capable d'analyser les données capitalisées pour enrichir les connaissances existantes, voir même pour en créer ;
- les raisonnements du moteur d'inférences sont explicitables clairement par la présentation des probabilités et de la structure de dépendances des différents éléments de la configuration du **PEN** ;
- un réseau bayésien peut supporter les données discrètes et continues. Certains algorithmes d'apprentissage sont spécifiquement adaptés au cas où la base d'apprentissage serait incomplète (une variable spécifique peut être rarement observée). Il est donc possible de traiter toutes les variables de la configuration d'un **PEN** et d'étudier des configurations de processus de **PEN** rarement essayées ;
- la base probabiliste des réseaux bayésiens rend le mode de raisonnement suffisamment souple pour traiter des méthodes à comportement aléatoires, comme certains algorithmes d'optimisation, ou certains type de **PEN** ;



### *CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES*

---

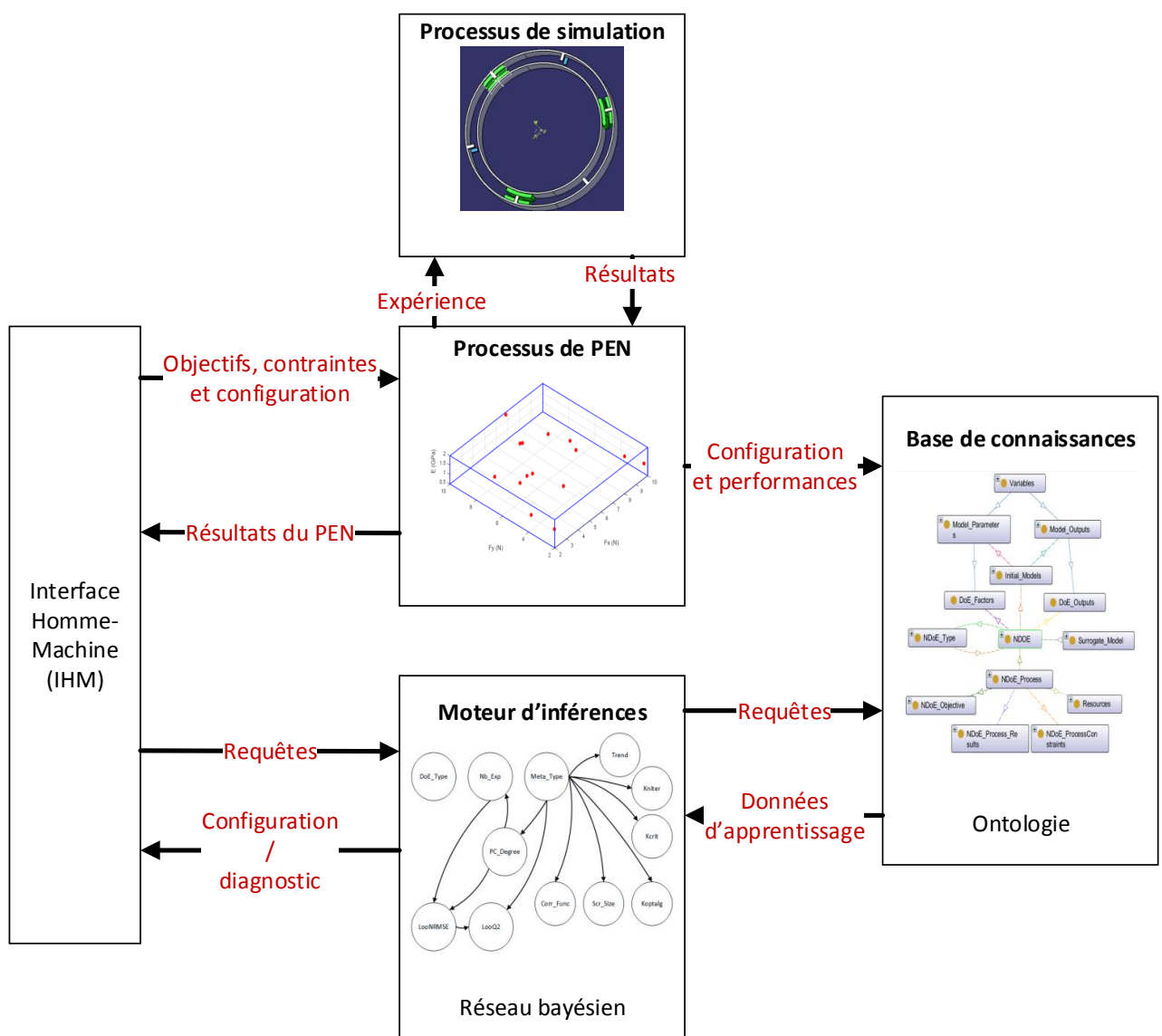
- l'utilisateur reçoit des prédictions de configurations de processus de **PEN** adaptées à ses exigences, ou un pronostic des performances d'une configuration donnée, toujours accompagnées d'indicateurs de performances prédictives du réseau bayésien lui-même. L'utilisateur sait ainsi si la phase d'apprentissage s'est correctement déroulée et s'il peut faire confiance aux prédictions du système à base de connaissances.

Cette application des réseaux bayésiens comble un manque de développements constatés lors de l'état de l'art (cf. chapitre 2) pour l'aide à la configuration des processus de **PEN** et conduira les entreprises à mettre en place plus rapidement des démarches de **PEN** au plus tôt dans le processus de développement de produits. Les méthodes existantes pour réduire le coût d'exécution d'un processus de **PEN** pourront être pleinement exploitées et rapidement mises en place. Le processus de **PEN** sera configuré de manière à fournir le plus d'informations possibles à partir d'un minimum d'expériences. Le comportement du produit pourra alors être prédit au plus tôt, tout en intégrant toutes les sources de variabilités liées à ce produit.

### 3.4 Synthèse et discussion

Le système à base de connaissances proposé est principalement basé sur une ontologie dédiée aux processus de PEN et sur un moteur d'inférences bayésien effectuant des prédictions sur la base des processus de PEN historisés (Figure 3.20). L'ontologie permet de capitaliser les connaissances de façon structurée et le réseau bayésien réutilise les données et les connaissances pour proposer de nouvelles configurations du processus de PEN pour un contexte donné.

FIGURE 3.20 – Architecture du système à base de connaissances proposé



### CHAPITRE 3. PROPOSITION : SYSTÈME À BASE DE CONNAISSANCES POUR L'AIDE À LA CONFIGURATION DES PROCESSUS DE PLAN D'EXPÉRIENCES NUMÉRIQUES

L'ontologie proposé, ODE, apporte une structuration partageable des connaissances liées aux processus de **PEN** tout en intégrant les développements existants. A partir de ses concepts clés, ODE peut être spécialisée suivant les ressources logiciels disponibles.

Le moteur d'inférences, basé sur les réseaux bayésiens, combine les connaissances expertes déjà acquises et la découverte de nouvelles connaissances à partir des données capitalisées à l'aide de ODE. Il offre des prédictions associées d'indicateurs de qualité. L'utilisateur dispose donc d'une assistance sur les choix qu'il va devoir faire pour configurer le processus de **PEN**, mais il sait également s'il peut faire confiance aux conseils donnés. Le réseau bayésien donne aussi une représentation claire et partageable de l'influence de chacun des choix sur les performances finales du processus de **PEN**. L'utilisateur peut alors se concentrer sur les choix essentiels.

Le lien entre ODE et le moteur d'inférences est assuré par le langage SPARQL. Ce langage permet d'exprimer des requêtes directement à l'ontologie. Ces requêtes, détaillées dans le chapitre 4, permettent de former une base d'apprentissage pour définir un réseau bayésien adapté.

Ce système est utilisable pour deux usages : identifier la configuration de **PEN** la plus probablement efficace ou estimer les performances probables d'une configuration donnée. Ce système à base de connaissances permettra, à terme, d'améliorer la rentabilité des processus de **PEN**, d'accélérer l'accès aux connaissances de l'entreprise, de les partager, en offrant une capitalisation structurée et une analyse automatique des connaissances. Il reste cependant à caractériser son efficacité et mener une campagne de validation, détaillée dans le chapitre suivant.

La principale piste d'amélioration est celle de la stratégie d'apprentissage. Beaucoup d'autres possibilités existent, notamment en termes d'apprentissage de la structure du réseau bayésien (nombreux algorithmes), de types de moteurs d'inférences (une hybridation réseau bayésien - réseau de neurones pourraient être envisagée) et de type d'apprentissage. Le seul type d'apprentissage considéré dans ces travaux est l'apprentissage supervisé (connaissance des processus de **PEN** et des résultats associés). L'apprentissage non-supervisé (résultat des **PEN** inconnus) et l'apprentissage par renforcement (récompense selon la performance du processus de **PEN**) pourrait enrichir le fonctionnement du moteur d'inférences).

Le chapitre suivant met en œuvre cette proposition dans un cas concret utilisé lors du projet SDM4DOE.



# Chapitre 4

## Application du système proposé

■ *Les chiffres, c'est pas une science exacte figurez-vous!* ■

---

Jean-Christophe Hembert (Karadoc),  
Kaamelott, Livre 5, écrit par A.Astier

### Sommaire

---

<b>4.1</b>	<b>Implémentation</b>	<b>133</b>
4.1.1	Outils utilisés	133
4.1.2	Modèles de simulation numérique utilisés	135
4.1.3	Instanciation de la base de connaissances	138
<b>4.2</b>	<b>Exemple d'utilisation</b>	<b>141</b>
4.2.1	Situation initiale	141
4.2.2	Extraction des connaissances capitalisées	144
4.2.3	Apprentissage du réseau bayésien	146
4.2.4	Estimation des performances prédictives du réseau bayésien	153
4.2.5	Inférences du réseau bayésien : prédiction de la performance de la configuration de processus de plan d'expériences numériques	156
4.2.6	Inférences du réseau bayésien : prédiction de configurations performantes de processus de plan d'expériences numériques	156
4.2.7	Vérification et exécution de la configuration proposée	160
<b>4.3</b>	<b>Conclusion de l'application</b>	<b>162</b>

---

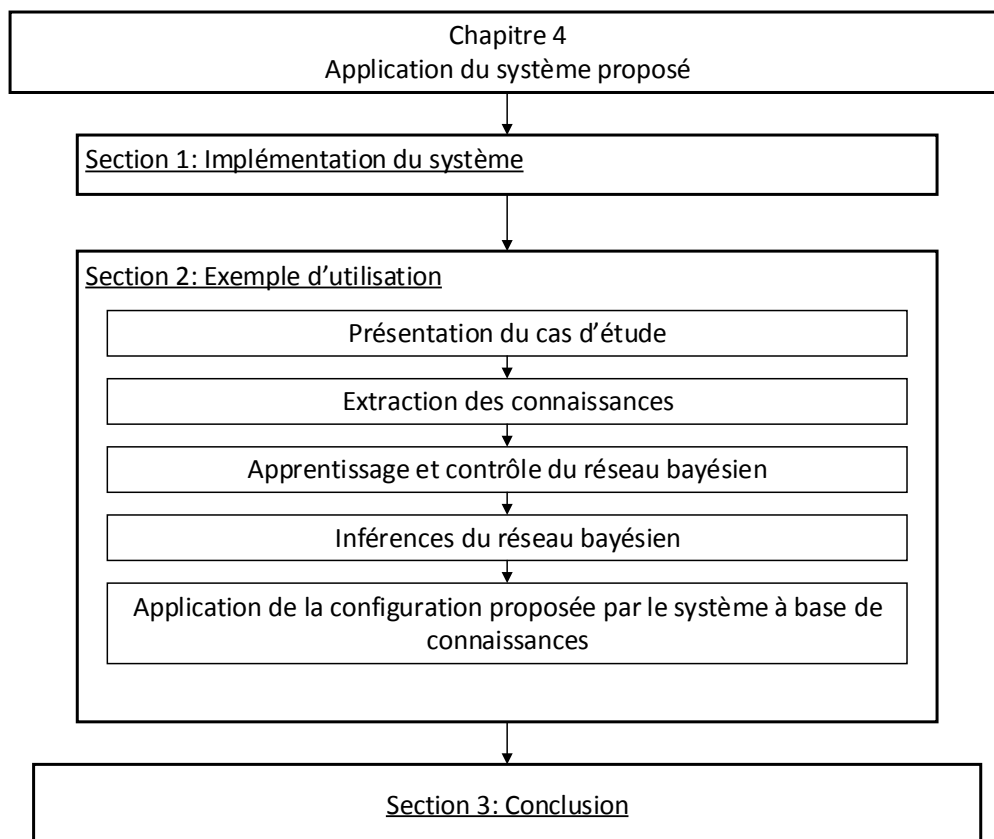


Le processus de **PEN** est un processus complexe à configurer, et un système à base de connaissances pour l'aide à la configuration a été proposé dans le chapitre 3. Ce système s'articule autour de deux éléments principaux :

- une gestion des connaissances, supportée par une ontologie spécifique au processus de **PEN**. Les connaissances peuvent alors être capitalisées et réutilisées à volonté ;
- un moteur d'inférences, permettant de proposer des configurations efficaces, compte-tenu des connaissances capitalisées et de la nature des besoins de l'utilisateur. L'état de l'art a conduit à choisir les réseaux bayésiens en tant que moteur d'inférences. Ce moteur exploite les connaissances capitalisées et peut prédire, soit les performances d'une configuration donnée, soit les configurations qui atteignent les performances désirées, pour un problème donné.

Le chapitre 4 (Figure 3.1) présente en premier lieu une implémentation du système proposé. Cette implémentation est utilisée dans la seconde section pour exposer un cas d'utilisation. En s'appuyant sur des modèles numériques utilisés dans l'industrie automobile, l'utilisation et les capacités du système à base de connaissances sont détaillées.

FIGURE 4.1 – Plan du chapitre 4





## 4.1 Implémentation

### 4.1.1 Outils utilisés

Le système à base de connaissances proposé est constitué de plusieurs éléments, déjà présentés au cours des chapitres précédents (Figure 4.2) :

1. la base de connaissances. L'ontologie, développée sur le logiciel Protégé 5, a été implémentée sur un serveur Virtuoso Universal Server<sup>®</sup>(OpenLink Software<sup>®</sup>). Les requêtes de données sont exécutées par l'utilisation du langage SPARQL ;
2. le moteur d'inférences. La mise en œuvre de réseaux bayésiens est assurée par le paquet R "bnlearn" ([Denis 14]). Cette bibliothèque de fonctions a été choisie notamment pour sa documentation riche et ses nombreux algorithmes d'apprentissages disponibles ;
3. une Interface Homme-Machine (IHM). L'IHM permet à l'utilisateur d'exploiter les données pertinentes à disposition, d'exécuter des processus de PEN et de communiquer avec le moteur d'inférences et la base de connaissances. Un modèle d'IHM est proposé à titre d'illustration pour le cas d'utilisation présenté tout au long de ce chapitre.
4. le processus de PEN. Le processus est exécuté par un ensemble de scripts, développés en langage Python et pilotant les différentes fonctions d'URANIE<sup>1</sup> (CEA). URANIE transmet les données nécessaires au solveur éléments-finis Code\_Aster<sup>2</sup> (EDF) pour exécuter les différentes simulations.

Le fonctionnement du système proposé est décrit par un logigramme (Figure 4.3), suivant la norme ISO 5807. Le système dispose de deux modes de fonctionnement. Dans les deux cas, les données de contexte (définissant le produit étudié, le modèle numérique utilisé, les facteurs et variables d'intérêt, les ressources à disposition, etc.) et les données stratégiques (objectif du processus de PEN, contraintes de l'étude) sont obligatoirement spécifiées par l'utilisateur (cf.chapitre 2). Les inférences réalisées par la suite se concentrent sur la configuration du processus de PEN. La configuration désigne tous les éléments nécessaires à l'exécution du processus PEN : le type de plan d'expériences, le nombre d'expériences, le type de modèle empirique, les paramètres internes associées à chaque méthode (ex : degré d'un modèle polynômial), etc.

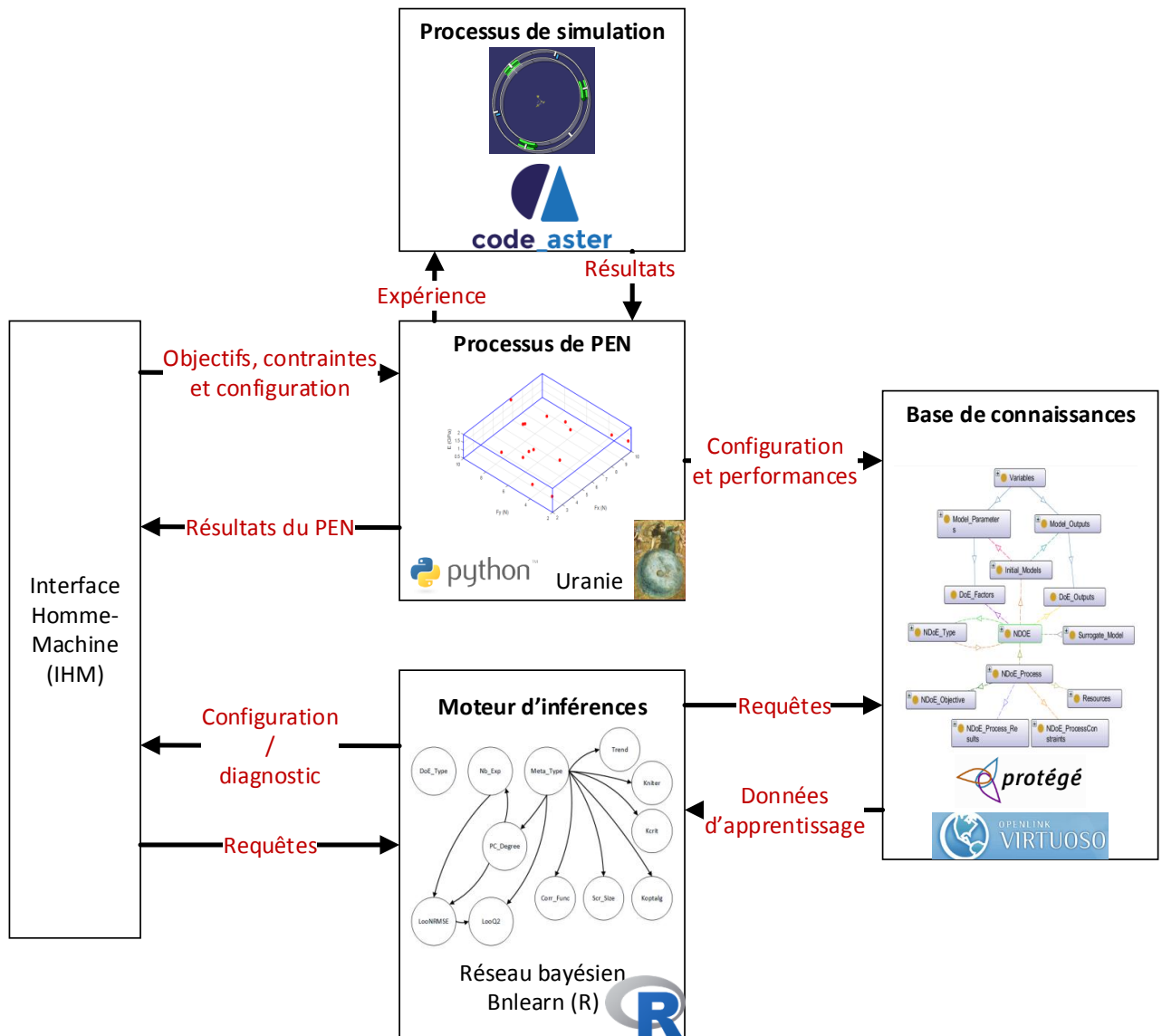
Le premier mode consiste à compléter la configuration du processus de PEN. Par rapport aux données spécifiées au préalable par l'utilisateur, les données concernant des cas similaires sont extraites de la base de connaissances. La notion de cas similaire est définie, dans le cadre de ces travaux, comme étant un processus de PEN se rapportant à un

---

1. <https://sourceforge.net/projects/uranie/>

2. <http://www.code-aster.org/>

FIGURE 4.2 – Outils utilisés pour le système à base de connaissances



même type de produit, à un même type de modèle numérique et de phénomène physique simulé, à un même type de démarche de PEN (adaptatif ou non) et à un même objectif (production d'un modèle de substitution, optimisation, analyse de sensibilité, etc.).

Le second mode est utilisé lorsque l'utilisateur dispose d'une configuration complète dont il ne connaît pas les performances. Le réseau bayésien estime ces performances (temps d'exécution, précision des résultats, etc.). Si les performances estimées ne sont pas conformes aux attentes de l'utilisateur, le système revient sur le premier mode de fonctionnement, afin de proposer une ou plusieurs autres configurations que le système estime conformes.

Chacun des deux modes de fonctionnement conduit à une configuration finale, validée par l'utilisateur. L'utilisateur est libre de modifier la configuration proposée par le système avant de la valider. La configuration est alors exécutée. Elle est capitalisée, avec les résultats du processus et ses performances mesurées, dans la base de connaissances.

Le processus de PEN considéré ici est détaillé en annexe (Annexe A), également sous forme de logigrammes.

### 4.1.2 Modèles de simulation numérique utilisés

Afin de mettre en place le système proposé, deux types de modèles ont été utilisés. Il s'agit de deux modélisations par la méthode des éléments finis de structures mécaniques, l'une étant un système théorique, l'autre étant un système issu du secteur automobile.

Le premier type de modèle est le type T. Les 6 modèles de type T, numérotés de T1 à T6, modélisent le comportement mécanique d'une structure en forme de T (Figure 4.4).

Le second type de modèle est le type Rotor. Les 2 modèles de types Rotor modélisent le comportement d'un système de découplage dynamique. Ce système est situé au sein du pulseur d'air, qui appartient au groupe moto-ventilateur d'un appareil de climatisation produit par Valeo (Figures 4.5 et 4.6). Le pulseur est constitué d'un rotor et d'un stator. La fonction principale du système de découplage est d'empêcher la transmission des vibrations liées à la rotation du rotor. Ce système de découplage est constitué de plusieurs plots en élastomère. Une démarche de plan d'expériences est mise en place sur ce modèle (Figure 4.7) afin d'identifier les facteurs influents sur la transmission des vibrations et d'optimiser le système pour réduire cette transmission.

Les caractéristiques principales des modèles T sont résumées dans la Table 4.1, et celles des modèles Rotor dans la Table 4.1.

FIGURE 4.3 – Logigramme du système à base de connaissances pour l'aide à la configuration du processus de PEN

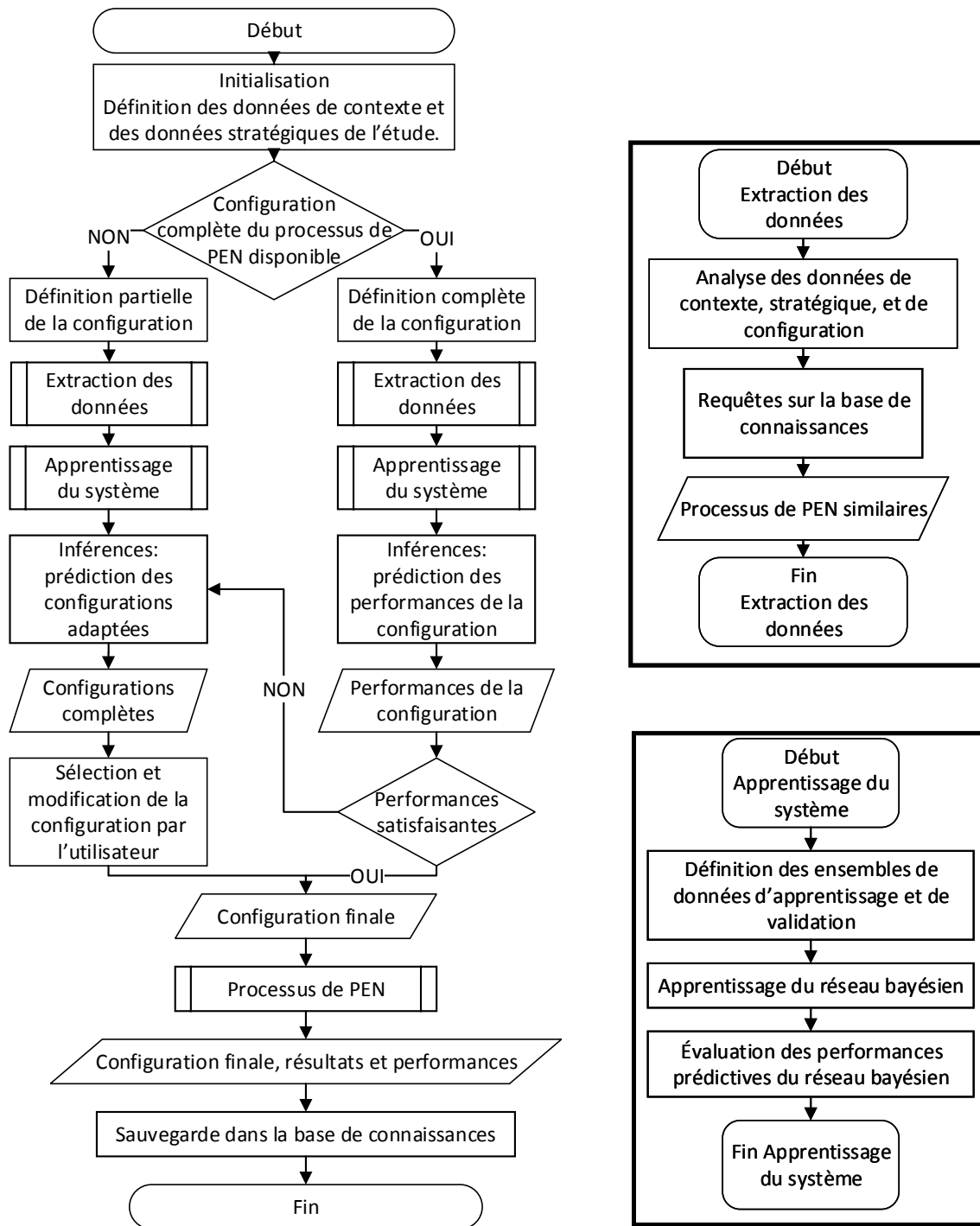


FIGURE 4.4 – Représentation 3D de la structure des modèles T

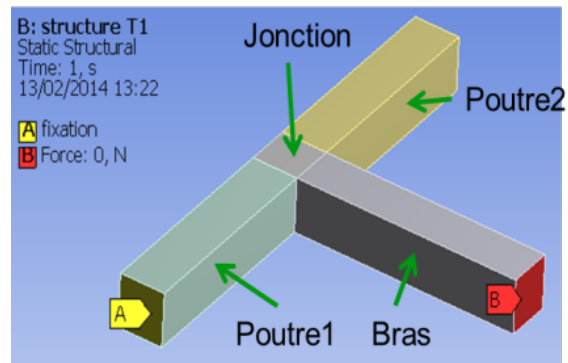


FIGURE 4.5 – Système de climatisation, comprenant le pulseur. Image Valeo.

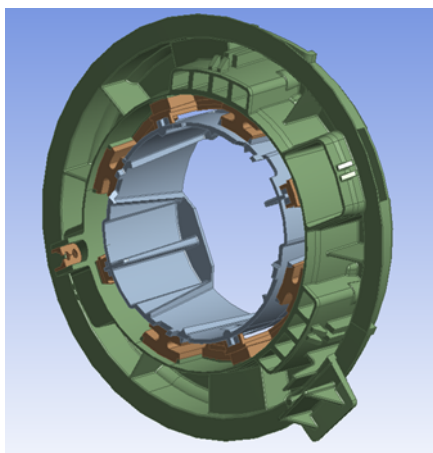


FIGURE 4.6 – Représentation 3D du système de découplage du pulseur

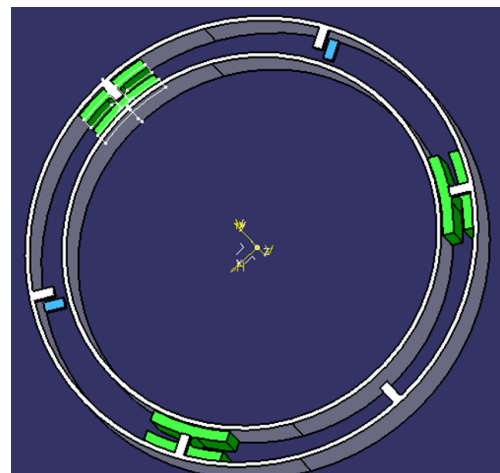


FIGURE 4.7 – Modélisation du système de découplage pour simulation

TABLEAU 4.1 – Caractéristiques principales des modèles numériques utilisés

Nom du modèle	Type de comportement	Facteurs	Réponses
T1	Statique linéaire	$E, F_x, F_y$	Reac ; Strain
T2	Statique linéaire avec contacts	$E, F_x, F_y$	Reac ; Strain
T3	Analyse modale - 3 matériaux différents	$E_1, E_2, E_3$	Freq1, Freq2, Freq3
T4	Statique linéaire - 3 matériaux différents	$E_1, E_2, E_3$	Strain, Stress, Depl (dans la jonction)
T5	Thermo-mécanique	$\alpha_{/z}, \lambda_{bras/z}, \lambda_{jonction/z}$	Strain, Stress, Depl (dans la jonction)
T6	Statique non-linéaire (hyper-élasticité + force mobile)	$E, F_x, F_y$	Reac, Strain
Rotor1a	Analyse modale	6 facteurs géométriques, $E, \Delta_{-\varphi}$	Freq1, Freq2, Freq3 et FactPZ1, FactPZ2, FactPZ3
Rotor1b	Analyse modale (amortissement+effet gyroscopique)	6 facteurs géométriques, $E, \Delta_{-\varphi}$	5 premières fréquences critiques (diagramme de Campbell)

Légende : **E** : Module d'Young ; **F<sub>x</sub>** : composante selon l'axe x de la force appliquée ; **Reac** : force résultante aux encastremets ; **Strain** : Maximum du premier invariant du tenseur des déformations ; **Stress** : Maximum du premier invariant du tenseur des contraintes ; **Depl** : déplacement maximal), **Freq** : fréquence propre ; **FactPZ** : facteur de participation du mode selon l'axe z ;  **$\alpha_{/z}$**  : Coefficient de dilatation thermique suivant l'axe z ;  **$\lambda_{bras/z}$**  : coefficient de conductivité thermique suivant l'axe z sur le bras ;  **$\lambda_{jonction/z}$**  : coefficient de conductivité thermique suivant l'axe z sur la jonction ;  **$\Delta_{-\varphi}$**  : déphasage des balourds

### 4.1.3 Instanciation de la base de connaissances

Afin de tester les capacités du système proposé, une base de connaissances a été constituée en générant des configurations de processus de PEN et en les exécutant, pour chaque modèle numérique. Ces données sont ensuite exploitées par le système à base de connaissances.

Pour chaque modèle, 2000 configurations de processus de PEN ont été générées, et 1773 pour les processus de PEN adaptatifs (Table 4.2). Les processus de PEN considérés ont tous pour objectif la métamodélisation. Cet objectif implique la définition de nombreux éléments de configuration. De plus, la production d'un modèle de substitution peut permettre de remplir, par la suite, les objectifs d'analyse de sensibilité et d'optimisation.

Tous ces cas ont été générés de façon aléatoire, suivant des lois uniformes. Toutes les modalités de chaque variable ont une probabilité identique d'apparaître. Quelques exceptions existent, telles que pour la variable décrivant le degré d'un modèle de chaos polynômial, qui doit respecter une règle de dépendance connue par rapport au nombre

TABLEAU 4.2 – Définition des ensembles de données générés pour la base de connaissances.

N° de l'ensemble de données	1	2	3	4	5	6	7	8	9
Modèle numérique	T1	T1	T2	T3	T4	T5	T6	R1a	R1b
Stratégie de planification d'expériences	S	A	S	S	S	S	S	S	S
Nombre de processus de PEN exécutés	2000	1773	2000	2000	2000	2000	2000	2000	2000

Légende : S : PEN statique ; A : PEN adaptatif

TABLEAU 4.3 – Variables et modalités prises en compte pour la génération des configurations de processus de PEN

Variables de configuration	Plans statiques
Type de plan	Halton, Sobol, hypercube latin
Nombre d'expériences	Nombre entier entre 10 et 50
Type de modèle empirique	Krigeage, chaos polynômial
Paramètres du modèle de krigeage	
Fonction de corrélation	Exponentiel, Gauss, Matern3/2, Matern5/2, Matern7/2
Fonction de tendance	Constante ou linéaire
Critère d'optimisation	Maximum de vraisemblance, Maximum de vraisemblance restreint, Leave-One-Out
Algorithme d'optimisation	NelderMead, BOBYQA, BFGS
Nombre d'itérations	T : 100 ; 500 ; 1000 ; 1500, Rotor : 100 ; 200 ; 500 ; 800 ; 1000 ; 1200 ; 1500
Taille de l'échantillon de screening	T : 0 ; 100 ; 250 ; 500, Rotor : 0 ; 100 ; 200 ; 300 ; 400 ; 500
Paramètres du modèle de chaos polynômial	
Degré	1, 2 ou 3 suivant le nombre d'expériences

minimal d'expériences (cf.chapitre 1). Les variables et modalités prises en compte pour la génération de ces configurations sont résumées dans les tables 4.3 et 4.4.

Une fois définies, toutes ces configurations ont été exécutées à l'aide des scripts Python développés, d'URANIE et de Code\_Aster. Chaque configuration, accompagnée de l'ensemble des données générées, forme une instance du processus de PEN. Chaque instance est intégrée à la base de connaissances selon la structure apportée par l'ontologie.

TABLEAU 4.4 – Variables et modalités prises en compte pour la génération des configurations de processus de PEN adaptatifs

<b>Variables de configuration</b>	<b>Plans adaptatifs</b>
Type de plan initial	Halton, Sobol, hypercube latin
Nombre d'expériences initial	Nombre entier entre 10 et 50
Type de modèle empirique	Krigeage
<b>Paramètres du modèle de krigeage</b>	
Fonction de corrélation	Exponentiel, Gauss, Matern 3/2, Matern 5/2, Matern 7/2
Fonction de tendance	Constante ou linéaire
Critère d'optimisation	Maximum de vraisemblance, Maximum de vraisemblance restreint, Leave-One-Out
Algorithme d'optimisation	NelderMead, BOBYQA, BFGS
Nombre d'itérations	T : 100;500;1000;1500, Rotor : 100;200;500;800;1000;1200;1500
Taille de l'échantillon de screening	T : 0;100;250;500, Rotor : 0;100;200;300;400;500
<b>Paramètres généraux des plans adaptatifs</b>	
Métaheuristique pour la recherche de nouvelles expériences	Essaim Particulaire (PSO), Algorithme Génétique (GA)
Nombre maximal d'expériences à ajouter	Nombre entier entre 5 et 20
Objectif de précision du modèle empirique (LooNRMSE_REF)	0.1 ; 0.01 ; 0.001
Objectif de prédictivité du modèle empirique (LooQ <sup>2</sup> _REF)	0.90 ; 0.95 ; 0.98
Critère de remplissage	Expected Improvement ; Variance maximale
Seuil de convergence des métaheuristiques	0.001 ; 0.0001 ; 0.00001
<b>Paramètres de l'algorithme génétique</b>	
Population	Nombre entier entre 5 et 20
Taux de mutation	0.01 ; 0.1 ; 0.2 ; 0.5 ; 0.8
Taux de survie	0.1 ; 0.3 ; 0.5 ; 0.8 ; 0.9
<b>Paramètres de l'algorithme d'essaim particulaire</b>	
Population	Nombre entier entre 10 et 80
Nombre maximal d'évaluation	100 ; 300 ; 500 ; 800 ; 1000
Nombre de particules	Nombre entier entre 5 et 30
Mémoire globale (Size)	10 ; 20 ; 30 ; 50 ; 100
nombre d'individus avant mise à jour de la mémoire (Step_Size)	Nombre entier entre 2 et 10



## 4.2 Exemple d'utilisation

### 4.2.1 Situation initiale

L'utilisation du système d'aide à la décision est présentée à l'aide d'un scénario spécifique. L'évolution de la situation est illustrée par les différentes fenêtres de l'IHM. Initialement, la base de connaissances contient toutes les instances définies dans la section précédente. Dans le cadre de cet exemple, l'utilisateur du système souhaite obtenir un modèle de substitution du modèle Rotor1a afin d'accélérer les prises de décisions et optimiser le système de pulseur d'air. Pour cela, il dispose déjà d'une configuration de processus de PEN complète, qu'il propose au système, mais dont il ne connaît pas les performances. Ce type de scénario permet d'illustrer l'ensemble des étapes décrites dans le logigramme globale (Figure 4.3).

L'utilisateur débute le processus de PEN par la définition des données de contexte (Figure 4.8). Il définit notamment l'identifiant du processus de PEN, les identifiants du produit et du projet auquel est rattaché le processus de PEN, le modèle de simulation numérique, les facteurs du PEN, etc.

FIGURE 4.8 – IHM : fenêtre d'initialisation du processus de PEN

**Contexte de l'étude**

Identifiant du projet:   Identifiant du processus de PEN:

Identifiant du produit:  Répertoire de travail:

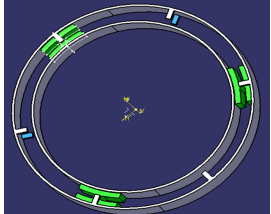
Date:  Auteur:

Modèle de simulation:

---

**Réponses du modèle**

Nom	Grandeur	Unité	Type	Loi	Min	Max
Epaisseur_patin	géométrie	m	Continue	Uniforme	0.0015	0.0045
longueur_patin	géométrie	m	Continue	Uniforme	0.01	0.0045
hauteur	géométrie	m	Continue	Uniforme	0.01	0.02



---

**Facteurs**

Nom	Grandeur	Unité
Freq1	Fréquence propre N°1	Hz
Freq2	Fréquence propre N°2	Hz
Freq3	Fréquence propre N°3	Hz
FactPZ1	facteur de participation du mode 1	N/A

Choix de variable d'intérêt

---

**Test unitaire**

Nombre de CPU:     Résultats valides

Temps de calcul estimé:  Durée du calcul de test:

L'utilisateur définit ensuite la configuration du processus de PEN qu'il propose dans l'IHM (Figure 4.9). Les éléments principaux de cette configuration sont détaillés dans le Tableau 4.5.

FIGURE 4.9 – IHM : fenêtre de configuration initial du processus de PEN

### Définition du processus de PEN

**Objectif Principal**

Métamodélisation

**Contraintes**

Durée maximale 1h

Finir au plus tôt

Finir pour: 15/09/2016 14h00

Plan d'expériences adaptatif

Précision minimale du métamodèle (Loo-NRMSE) 0.05

Prédictivité minimale du métamodèle (Loo-Q<sup>2</sup>) 0.98

---

### Configuration du processus

**PEN**

Type de démarche Statique

Type de plan Plan de Halton Nombre d'expériences 15

**Modèle mathématique**

Type de modèle Krigeage Fonction de corrélation Matern 7/2 Algorithme d'optimisation interne BOBYQA Nombre d'itérations 800

Fonction de tendance linéaire Critère d'optimisation Leave\_One\_Out Taille du screening initial 200

Le processus de PEN va produire un modèle caractérisé par la précision de ses résultats par rapport à ceux du modèle initial ("Rotor1a"), et sa prédictivité.  
 La précision est quantifiée par l'indicateur LooNRMSE: "Normalized Root Mean Square"; obtenue par la méthode de validation croisée "Leave-One-Out".  
 La précision est quantifiée par l'indicateur LooQ<sup>2</sup>, et doit être le plus proche possible de 1.

---

**Configuration complète détectée.**

Cliquez sur "Evaluer" pour obtenir une estimation de la précision et de la prédictivité du modèle obtenue par cette configuration du processus de PEN

**Evaluer**

TABLEAU 4.5 – Caractéristiques principales de la configuration complète du processus de PEN proposée par l'utilisateur

<b>Caractéristiques de configuration</b>	<b>Valeur</b>
Objectif	production d'un modèle de substitution
Précision minimale exigée (LooNRMSE)	0.05
Prédictivité minimale exigée (LooQ <sup>2</sup> )	0.98
Stratégie de planification d'expériences	Statique
Modèle numérique	Rotor1a
Type de plan initial	Halton
Nombre d'expériences	15
Type de modèle empirique	Krigeage
<b>Paramètres du modèle de krigeage</b>	
Fonction de corrélation	Matern 7/2
Fonction de tendance	linéaire
Critère d'optimisation	Leave-One-Out
Algorithme d'optimisation	BOBYQA
Nombre d'itérations	800
Taille de l'échantillon de screening	200

## 4.2.2 Extraction des connaissances capitalisées

A partir de cette configuration proposée, le système va rechercher au sein de la base de connaissances les processus de **PEN** similaires au cas traité. Cette similarité est évaluée suivant l'objectif du processus de **PEN**, le type de modèle numérique considéré et le type de démarche de planification des expériences. Cette stratégie d'extraction a été choisie pour mettre en application l'approche "multi-net" choisie au chapitre 3. Cette approche consiste à réaliser un réseau bayésien par objectif, type de modèle numérique et type de démarche de planification d'expériences. L'objectif est de constituer des réseaux bayésiens simples et performants.

L'ontologie est instanciée avec toutes les configurations de processus de **PEN** générées, par le biais de Virtuoso Universal Server®. La structure de l'ontologie est définie dans le modèle "TBox" (T : Terminologique) et reste séparée des instances, définies quant à elles dans la "ABox" (A :Assertion). Ces deux parties sont intégrées dans le serveur pour y être exploitées.

La première requête (Extrait de codes 4.1) extrait les processus de **PEN** ayant un objectif de métamodélisation portant sur le modèle Rotor1a (Figure 4.1). Les résultats de cette requête sont présentés dans la Table 4.6.

Extrait de codes 4.1 – Requête SPARQL N°1

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ode: <http://www.semanticweb.org/gblondet/ontologies/2016/ODE#>

SELECT ?NDOE_Process ?obj_type ?Model_Name ?Surrogate_Model
FROM <http://localhost.utc.fr:8890/ODE/ABOX>
WHERE
{{
  ?NDOE_Process ode:hasObjective ?objective .
    ?objective rdf:type ?obj_type .
  ?NDOE_Process ode:hasNDoE ?NDOE .
  ?Surrogate_Model ode:IsTrainedBy ?NDOE .
  ?NDOE ode:hasInitModel ?INIT_MODEL .
  ?INIT_MODEL ode:ModelName ?Model_Name .
  FILTER regex(str(?Surrogate_Model), ".^")
  FILTER (?obj_type!= owl:NamedIndividual
    && ?obj_type!= ode:NDoE_Objective)
}}

```

Cette première requête a permis la sélection des deux mille configurations correspondantes au modèle "Rotor1a" et à un objectif de métamodélisation. Dans le cas présent, l'extraction des processus de **PEN** adaptatifs est inutile, sachant qu'aucun processus adap-

TABLEAU 4.6 – Extraction des données capitalisées par la requête n°1

<b>NDOE_Process</b>	<b>obj_type</b>	<b>Model_Name</b>	<b>Surrogate_Model</b>
#NDOE_Process_1	#Surrogate_Modelling	"Rotor1a"	#Surr_model_1
#NDOE_Process_10	#Surrogate_Modelling	"Rotor1a"	#Surr_model_10
#NDOE_Process_2	#Surrogate_Modelling	"Rotor1a"	#Surr_model_2
#NDOE_Process_3	#Surrogate_Modelling	"Rotor1a"	#Surr_model_3
#NDOE_Process_4	#Surrogate_Modelling	"Rotor1a"	#Surr_model_4
#NDOE_Process_5	#Surrogate_Modelling	"Rotor1a"	#Surr_model_5
#NDOE_Process_6	#Surrogate_Modelling	"Rotor1a"	#Surr_model_6
#NDOE_Process_7	#Surrogate_Modelling	"Rotor1a"	#Surr_model_7
#NDOE_Process_8	#Surrogate_Modelling	"Rotor1a"	#Surr_model_8
#NDOE_Process_9	#Surrogate_Modelling	"Rotor1a"	#Surr_model_9

tatif n'a été généré pour ce modèle numérique.

La seconde requête permet d'extraire certains détails des configurations obtenues par la première requête (Extrait de codes 4.2). Les résultats sont présentés par la Table 4.7. Cette requête permet au système d'accéder à la valeur de plusieurs variables de configuration. D'autres requêtes, de la même forme que celle présentée, sont exécutées pour obtenir tous les détails des configurations, mais également les performances de chacune d'elles.

TABLEAU 4.7 – Extraction des données capitalisées par la requête n°2 (10 premières données)

<b>NDOE_Process</b>	<b>Nb_Exp</b>	<b>DoE_Type</b>	<b>Surrogate_Model_Type</b>
#NDOE_Process_1	"45"	#LatinHypercube	#Kriging
#NDOE_Process_10	"22"	#Halton	#PolynomialChaos
#NDOE_Process_2	"23"	#Halton	#PolynomialChaos
#NDOE_Process_3	"25"	#LatinHypercube	#Kriging
#NDOE_Process_4	"40"	#Sobol	#PolynomialChaos
#NDOE_Process_5	"15"	#Sobol	#PolynomialChaos
#NDOE_Process_6	"23"	#Halton	#Kriging
#NDOE_Process_7	"25"	#Sobol	#PolynomialChaos
#NDOE_Process_8	"50"	#Halton	#PolynomialChaos
#NDOE_Process_9	"34"	#Sobol	#PolynomialChaos

A l'issue de cette étape d'extraction de données de la base de connaissances, un ensemble de données est formé. Chaque donnée est constituée de la configuration détaillée de tous les processus de PEN pertinents, ainsi que des performances de ces processus. L'ensemble des données extraites est alors utilisé pour l'étape d'apprentissage automatique du réseau bayésien.

Extrait de codes 4.2 – Requête SPARQL N°2

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ode: <http://www.semanticweb.org/gblondet/ontologies/2016/ODE#>

SELECT ?NDOE_Process ?Nb_Exp ?DoE_Type ?Surrogate_Model_Type
FROM <http://localhost.utc.fr:8890/ODE/ABOX>
WHERE
{{
  ?NDOE_Process ode:hasObjective ?objective .
    ?objective rdf:type ?obj_type .
  ?NDOE_Process ode:hasNDoE ?NDOE .
  ?Surrogate_Model ode:IsTrainedBy ?NDOE .
  ?NDOE ode:hasInitModel ?INIT_MODEL .
  ?INIT_MODEL ode:ModelName ?Model_Name .
  ?NDOE ode:hasNDoEType ?NType .
  ?NDOE ode:NumberOfExperiments ?Nb_Exp .
  ?NType rdf:type ?DoE_Type .
  ?Surrogate_Model ode:SurMod_hasType ?SurrMod_Type .
  ?SurrMod_Type rdf:type ?Surrogate_Model_Type .
  FILTER regex( str(?Surrogate_Model), "^.")
  FILTER (?DoE_Type!= owl:NamedIndividual
    && ?DoE_Type!= ode:NDoE_Type )
  FILTER (?Surrogate_Model_Type!= owl:NamedIndividual
    && ?Surrogate_Model_Type!= ode:NDoEModels
    && ?Surrogate_Model_Type!= ode:Surrogate_Model
    && ?Surrogate_Model_Type!= ode:Surrogate_ModelType)
}}

```

### 4.2.3 Apprentissage du réseau bayésien

L'ensemble des données extraites de l'ontologie, lors de l'étape précédente, est utilisé pour la phase d'apprentissage du réseau bayésien. Cet ensemble est partitionné en deux sous-ensembles distincts, afin de réaliser une validation croisée de l'apprentissage. La validation croisée consiste à s'assurer, une fois l'apprentissage terminé, que le réseau bayésien dispose de bonnes performances prédictives et que l'on peut faire confiance en ses prédictions. Le premier sous-ensemble est utilisé pour l'apprentissage de la structure et des paramètres du réseau bayésien. Il contient 80% des données extraites. Le second, l'ensemble de validation, est utilisé pour évaluer les capacités de généralisation du réseau bayésien, et contient 20% des données extraites. Les données de chaque ensemble ont été choisies aléatoirement parmi toutes les données extraites. Le choix de ces proportions et des données intégrées dans l'une ou l'autre base est arbitraire, mais reste un choix courant ([Dreyfus 11]). Une seconde démarche de validation croisée, utilisant plusieurs partitionnements différents, est mise en œuvre lors de la phase d'évaluation des performances prédictives du réseau bayésien.

Cette seconde phase vise à évaluer la sensibilité de la validation croisée à la définition des partitions.

Une fois les deux sous-ensembles de données définis, le réseau bayésien est déterminé à partir des données extraites de la base de connaissances. Les variables et leurs modalités sont définies au préalable. Cette définition est représentée par la Figure 4.10 et la Figure 4.11. Ces deux histogrammes permettent de visualiser :

- l'ensemble des variables prises en compte dans le réseau bayésien ;
- l'ensemble des modalités de chacune de ces variables ;
- la composition de l'ensemble d'apprentissage suivant chacune des modalités de chaque variables (à gauche). On peut constater que l'ensemble d'apprentissage n'est pas équilibré. Certaines modalités sont plus représentées que d'autres pour une même variable ;
- la composition de l'ensemble de validation (à droite). Il est alors possible de comparer les répartitions entre les deux ensembles. Pour certaines modalités, il y a moins de données d'apprentissage que de validation. Dans ce cas, on peut s'attendre à un apprentissage de moins bonne qualité.

FIGURE 4.10 – Composition des ensembles d'apprentissage et de validation (1/2)

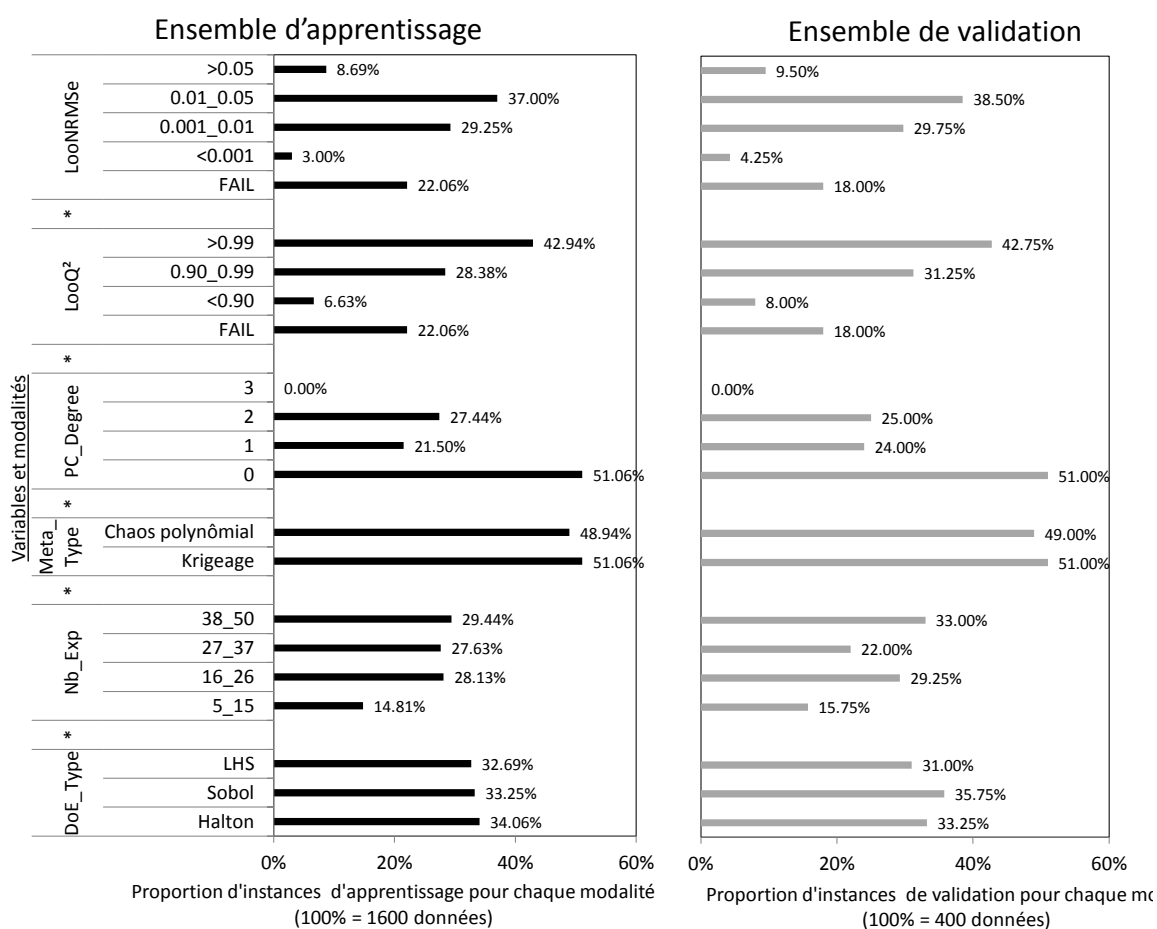
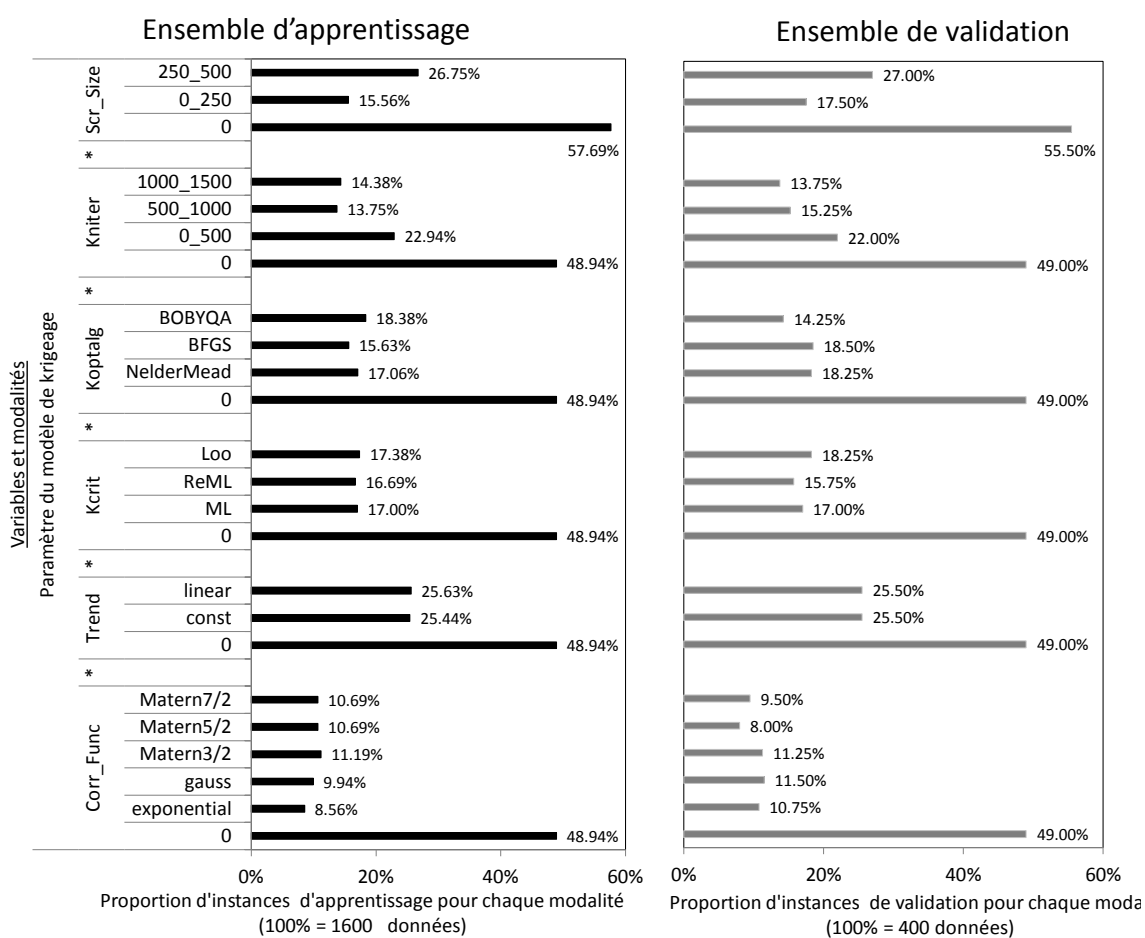


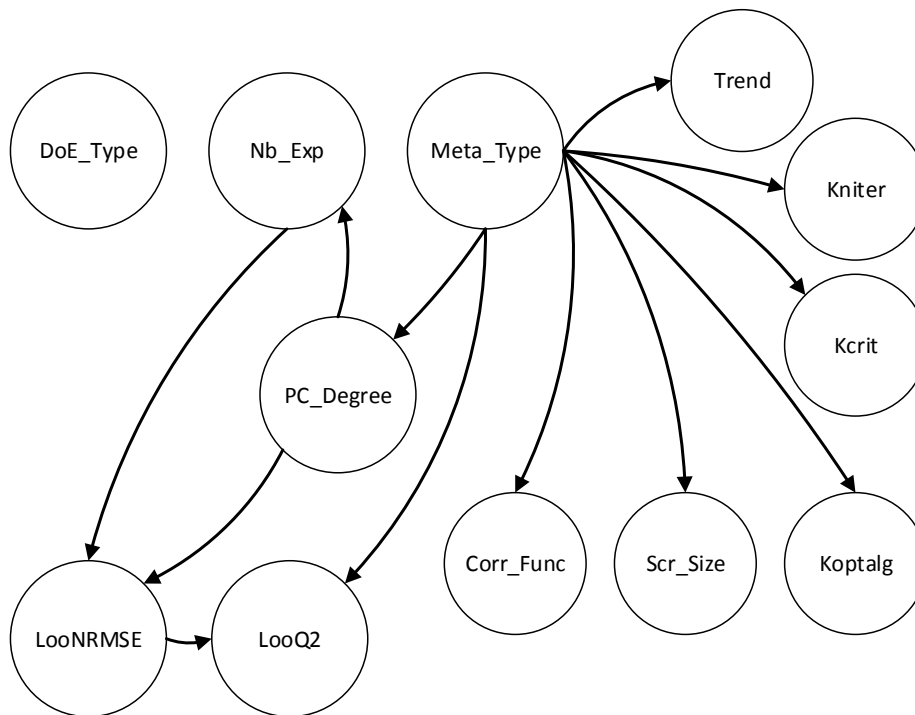


FIGURE 4.11 – Composition des ensembles d'apprentissage et de validation (2/2)



L'apprentissage de la structure est assurée par l'algorithme Hill-Climbing ([Denis 14]) fourni par bnlearn, permettant d'obtenir directement un graphe acyclique dirigé. Au préalable, certaines connaissances expertes peuvent être spécifiées, bien qu'elles ne soient pas obligatoires. Par exemple, certains arcs peuvent être interdits, d'autres peuvent être imposés. La structure résultante de l'apprentissage, ici sans introduction de connaissances sur les arcs du graphe, est illustrée par la Figure 4.12.

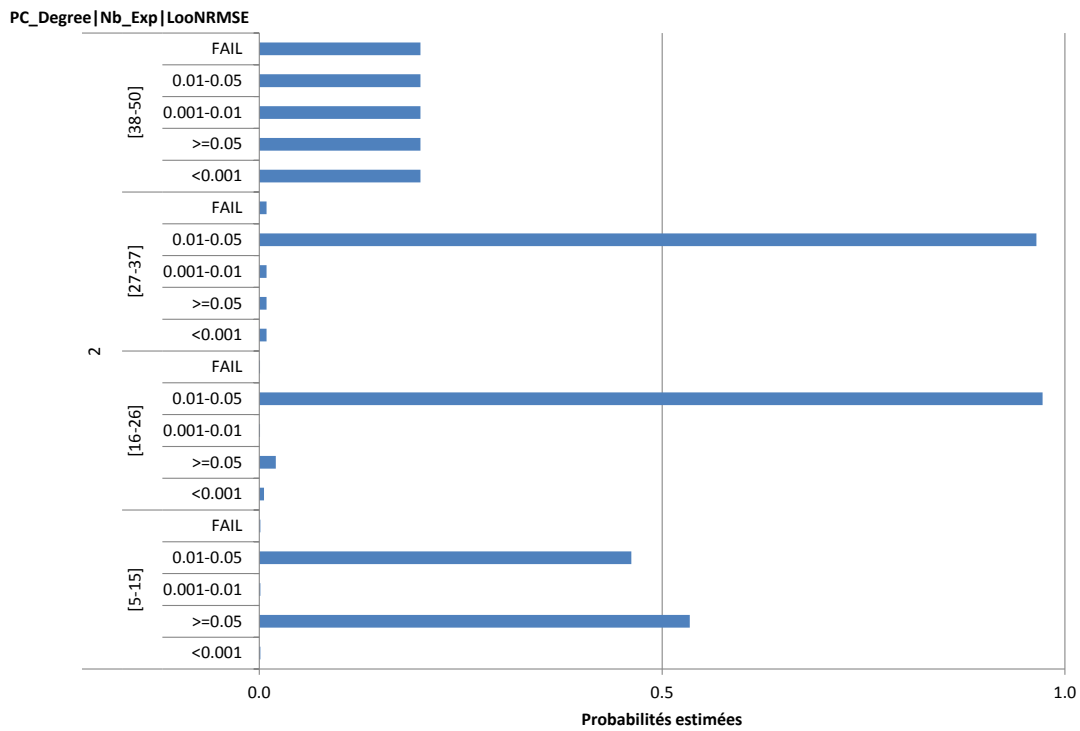
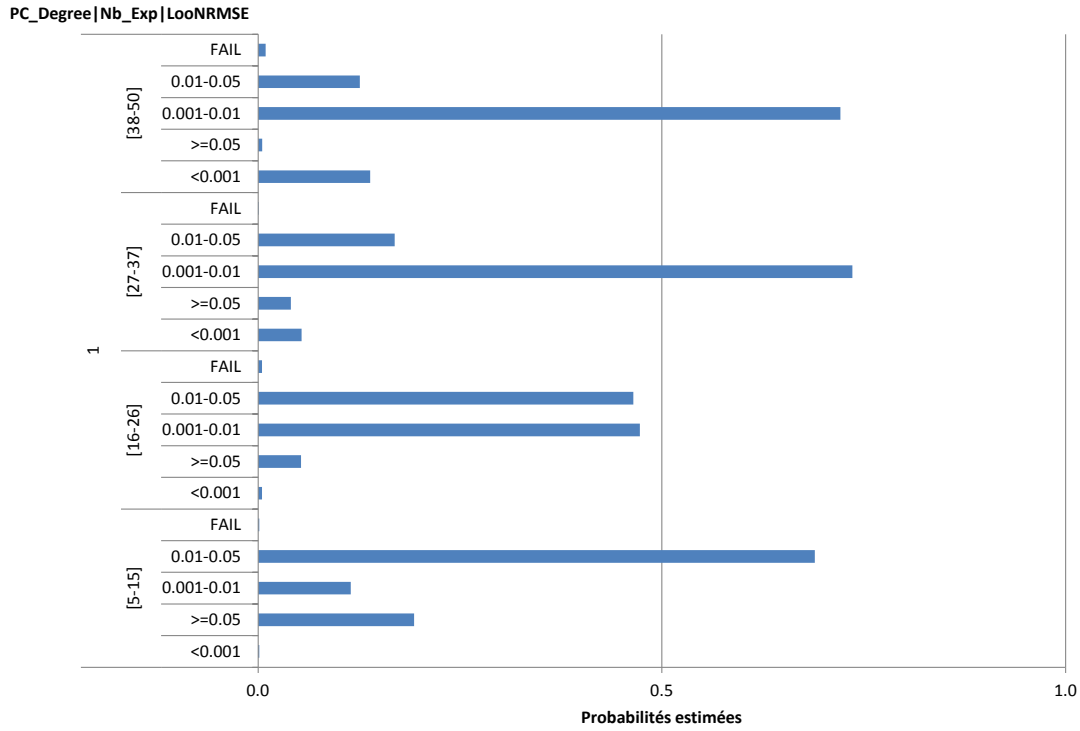
FIGURE 4.12 – Graphe du réseau bayésien après apprentissage sur 1600 configurations



Cette structure montre explicitement les dépendances entre les différentes variables considérées. Le type de plan n'a aucune influence, dans ce cas, sur le reste de la configuration. Une dépendance entre le degré du chaos polynômial et le nombre d'expériences a été identifiée. Cette dépendance était attendue, car elle a été imposée lors de la génération des données d'apprentissage, suivant la loi formulée au chapitre 1. Le choix du type de modèle de substitution a une influence sur la précision et la prédictivité. La variable "PC\_Degree" influe sur LooNRMSE et LooQ2. Cela signifie, d'après la définition des modalités (Figures 4.11 et 4.10), que le choix du krigeage (qui est équivalent, dans notre cas, à PC\_Degree=0) ou du chaos polynômial (PC\_Degree ≠ 0) impacte directement les performances du modèle de substitution. Le nombre d'expériences a une influence seulement sur LooNRMSE, et non sur LooQ2. En revanche, LooQ2 dépend de LooNRMSE, ce qui pourrait refléter le dilemme biais-variance.

Une fois la structure obtenue, les paramètres du réseau bayésien sont estimés. Cette estimation est réalisée ici par inférence bayésienne, réglée de façon à largement privilégier les données par rapport au prior bayésien uniforme (cf. Chapitre 3). Un exemple de résultat est présenté pour la variable LooNRMSE (Figure 4.13). Ces deux histogrammes représentent les probabilités conditionnelles estimées. Par exemple, la probabilité que LooNRMSE soit inférieure à 0.001 sachant que le nombre d'expériences est entre 38 et 50 et qu'un modèle de chaos polynômial de degré 1 est utilisé, est de 0.14.

FIGURE 4.13 – Probabilités conditionnelles estimées par inférence bayésienne. Exemple de la variable LooNRMSE, quantifiant la précision d'un modèle de substitution



#### 4.2.4 Estimation des performances prédictives du réseau bayésien

Une fois l'apprentissage réalisé, les techniques de validation croisée sont mises en œuvre. L'objectif est de donner à l'utilisateur des indicateurs lui permettant de vérifier si le réseau bayésien est capable de réaliser des prédictions fiables.

Comme détaillé dans le chapitre 3, l'ensemble de validation est utilisé pour obtenir une matrice de confusion par variable de configuration (Table 4.8). Les indicateurs par variable qui découlent de cette matrice sont, par exemple :

- TPR (taux de vrais positifs) : la probabilité que LooQ<sup>2</sup> soit prédit " $<0.90$ " et que c'est effectivement le cas est de 0.9688. On peut donc avoir raisonnablement confiance dans les prédictions de la modalité " $<0.90$ ";
- PPV (taux de prédictions positives) : 81.58% des prédictions sont réellement positives pour la modalité " $<0.90$ ";
- NPV (taux de prédictions négatives) : 99.72% des prédictions sont réellement négatives pour la modalité " $<0.90$ ".

Ces indicateurs permettent de quantifier le niveau de certitude des prédictions du réseau bayésien. Si TPR est à 100% pour une modalité donnée, les prédictions des processus de **PEN** dans cette modalité seront certainement correctes. Deux indicateurs globaux sont ajoutés :

- la mesure de précision globale pour chaque variable, obtenue à partir de la matrice de confusion. Pour la variable LooQ<sup>2</sup>, la précision globale est de 90.5% ;
- l'erreur moyenne globale de classification, obtenue par la méthode de validation croisée "k-fold". Associé à son écart-type, il donne une mesure de la sensibilité de l'erreur par rapport à l'ensemble de validation utilisé. Cette méthode a été appliquée avec  $k=5$  (5 partitions de l'ensemble des données), de manière à garder 400 données de validations et 1600 pour l'apprentissage, et  $n=100$  (100 partitionnements différents). l'erreur moyenne est de 11.27%, avec un écart-type de 0.0006. Il s'agit d'une erreur faible et peu sensible au choix des ensembles de validation et d'apprentissage.

Un modèle prédictif idéal aurait ses indicateurs TPR, TNR, PPV, NPV et Accuracy à 100% et les indicateurs FPR et l'erreur moyenne globale de classification à 0%. Cela est toutefois à modérer, car dans ce cas idéal, on pourrait suspecter un sur-apprentissage.

FIGURE 4.14 – IHM : fenêtre de rapport des performances prédictives du réseau bayésien

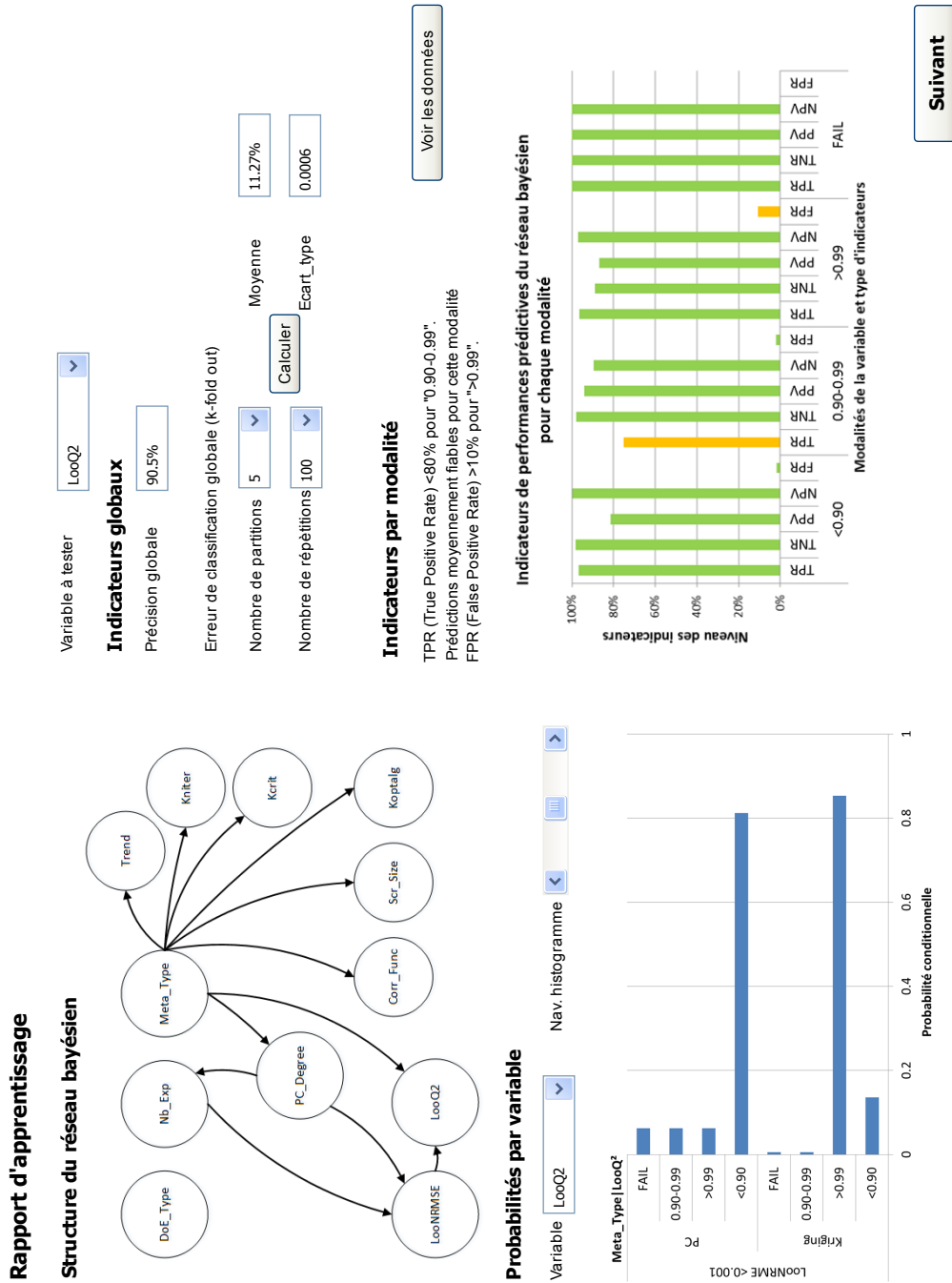


TABLEAU 4.8 – Matrice de confusion et indicateurs de performance prédictive du réseau bayésien, obtenus pour la variable LooQ<sup>2</sup> (prédictivité du modèle de substitution) après un apprentissage sur 1600 données à partir de la structure précédemment obtenue.

LooQ <sup>2</sup>	Modalités	Prédictions			
		<0.90	0.90-0.99	>0.99	FAIL
Observations	<0.90	31	0	1	0
	0.90-0.99	7	94	24	0
	>0.99	0	6	165	0
	FAIL	0	0	0	72
<b>Données</b>					
Nombre de données d'apprentissage		1600			
Nombre de données de validation		400			
<b>Indicateurs par modalités</b>					
Répartition des données d'apprentissage (/1600)		6.63%	28.38%	42.94%	22.06%
Répartition des données de validation (/400)		8.00%	31.25%	42.75%	18.00%
Nombre de vrais positifs (VP)		31	94	165	72
Nombre de faux positifs (FP)		7	6	25	0
Nombre de vrais négatifs (VN)		361	269	204	328
Nombre de faux négatifs (FN)		1	31	6	0
TPR ("Recall")		96.88%	75.20%	96.49%	100%
TNR ("Specificity")		98.10%	97.82%	89.08%	100%
FPR		1.90%	2.18%	10.92%	0.00%
PPV ("Precision")		81.58%	94.00%	86.84%	100%
NPV		99.72%	89.67%	97.14%	100%
<b>Indicateurs globaux</b>					
Accuracy*		90.5%			
Erreur moyenne globale de classification (k-fold)		11.27% (écart-type :0.0006)			

L'ensemble des informations nécessaires à l'utilisateur afin de valider l'apprentissage du réseau bayésien lui est affiché dans une fenêtre spécifique de l'IHM (Figure 4.14). Cet affichage permet à l'utilisateur de visualiser la structure du réseau, ainsi que les probabilités conditionnelles pour chaque variable. Il lui permet également de visualiser les performances prédictives du réseau bayésien pour chaque variable et modalité. L'affichage des indicateurs sous forme d'histogramme dynamique montre directement si un indicateur a une valeur anormale. Des explications sont également fournies à l'utilisateur pour une compréhension plus rapide du problème. L'utilisateur a ainsi la capacité de savoir s'il peut faire confiance aux prédictions du réseau bayésien, et ce pour chaque variable de la configuration du processus de PEN. L'apprentissage du réseau bayésien peut être contrôlé et validé si tous les indicateurs ont un niveau satisfaisant.

### 4.2.5 Inférences du réseau bayésien : prédiction de la performance de la configuration de processus de plan d'expériences numériques

Dans le cas où le réseau bayésien est validé, il est utilisé pour prédire les performances de la configuration pré-établie par l'utilisateur (Table 4.5). Il s'agit de classer la configuration selon les deux variables LooNRMSE et LooQ<sup>2</sup>. Cela revient à calculer les probabilités que cette configuration ait une performance donnée, sachant que la configuration est totalement connue. L'extrait de codes 4.3 donne la requête en langage R.

Extrait de codes 4.3 – Requête R n°1 : prédiction de performance d'une configuration de processus de PEN

```
cpquery(fitted ,
        event = (LooNRMSE == "0.01-0.05" & LooQ2 == "0.90-0.99"),
        evidence=(Nb_Exp == '[5-15]'
        & DoE_Type == "halton"
        & Meta_Type == "Krigage"
        & Corr_Func == "matern7/2"
        & Scr_Size == "]0;250]"
        & Kniter == "]500-1000]"
        & Trend == "linear" & Kcrit == "LOO"
        & Koptalg == "BOBYQA"))
```

Dans cet exemple, le résultat de cette requête est 0. Il est impossible, d'après le réseau bayésien utilisé ici, d'obtenir un modèle de substitution suffisamment précis et prédictif avec la configuration imposée. Il est alors très probable que cette configuration donne de mauvais résultats, compte-tenu des indicateurs de performances du réseau bayésien lui-même. L'utilisateur est invité à lancer la seconde phase, consistant à prédire de nouvelles configurations de processus PEN pour atteindre les objectifs fixés (Figure 4.15).

### 4.2.6 Inférences du réseau bayésien : prédiction de configurations performantes de processus de plan d'expériences numériques

Les performances du modèle de substitution produit par le processus de PEN n'étant pas satisfaisantes, le système d'aide à la décision cherche alors à proposer d'autres configurations, qui, elles, produiront un modèle de substitution plus performant. Tous les éléments de la configuration précédente sont ignorés. Seules les données de contexte et les données stratégiques sont conservées.



FIGURE 4.15 – IHM : fenêtre de prédiction des performances d’une configuration imposée

**Prédiction des performances du processus de PEN selon la configuration proposée par l'utilisateur**

**Rappel de la configuration utilisateur**

Objectif	Type de PEN	Nombre d'expériences initiales	Type de démarche	Type de métamodèle	Degré de la base polynômiale
Métamodélisation	LHS	15	PEN statique	Chaos polynômial	2

**Choix des variables objectifs et niveaux à atteindre**

Variables cibles      Modalités souhaitées

LooQ2      "0.90-0.99"     

LooNRMSE      "0.01-0.05"

     Probabilité estimée

La probabilité d'atteindre les objectifs fixés (LooQ<sup>2</sup> à "0.90-0.99" et LooNRMSE à "0.01-0.05") est **nulle**. Il est donc impossible d'atteindre ces objectifs avec cette configuration. Il est recommandé de prédire la configuration la plus adaptée selon les objectifs fixés.

Les étapes de recherches de cas similaires et d'apprentissage du réseau bayésien ont déjà été effectuées. La première extraction de données a été réalisée selon les données de contexte, selon le modèle numérique initial et selon le type de démarche de planification d'expériences numériques (ici, démarche non-adaptative). Le réseau bayésien utilisé précédemment est donc réutilisable directement. Ce réseau bayésien est cette fois-ci utilisé en sens inverse. Les modalités des variables "LooNRMSE" et "LooQ<sup>2</sup>" sont imposées, et la configuration est déduite.

La seconde requête (Extrait de codes 4.4) est alors évaluée sur toutes les configurations possibles (selon les modalités considérées dans le Tableau 4.3). Ainsi, le réseau bayésien a la capacité de proposer des configurations inédites à l'utilisateur. Cette requête produit une estimation des probabilités conditionnelles  $P(\text{LooNRMSE} == "0.01-0.05" \text{ et } \text{LooQ}^2 == "0.90-0.99" \mid \text{configuration})$ .

Extrait de codes 4.4 – Requête R n°2 : prédiction de configurations de processus de PEN

```
Pr[i] <- cpquery(fitted2, (LooNRMSE == "0.01-0.05"
& LooQ2 == "0.90-0.99"),
evidence = eval(parse(text=evi)))
```

Dans cet exemple, le système propose de nombreuses configurations atteignant certainement des objectifs fixés. Les réponses sont triées de façon à faire apparaître les configurations ayant la plus forte probabilité d’atteindre les objectifs fixés et les configurations exigeant le moins d’expériences possibles. Un extrait est présenté par le Tableau 4.16, présent dans l’IHM (Figure 4.17).

FIGURE 4.16 – Liste des configurations proposées par le système d’aide à la décision

	Nb_Exp	DoE_Type	Meta_Type	PC_Degree	Corr_Func	Scr_Size	Kniter	Trend	Kcrit	Koptalg	Pr
109	[5-15]	halton	Kriging	0	matern7/2	]0-250]	]1000-1500]	linear	ML	BOBYQA	1.0000000
156	[5-15]	sobol	Kriging	0	matern5/2	]250-500]	]0-500]	const	LOO	NelderMead	1.0000000
634	[5-15]	halton	Kriging	0	exponential	]250-500]	]1000-1500]	const	ML	BOBYQA	1.0000000
787	[5-15]	halton	Kriging	0	matern7/2	]250-500]	]500-1000]	linear	ML	NelderMead	1.0000000
794	[5-15]	halton	Kriging	0	matern7/2	]250-500]	]0-500]	linear	LOO	BOBYQA	1.0000000
902	[5-15]	halton	Kriging	0	gauss	]250-500]	]500-1000]	linear	LOO	BOBYQA	1.0000000
1031	[5-15]	halton	Kriging	0	matern3/2	]250-500]	]500-1000]	const	ReML	BFGS	1.0000000
1092	[5-15]	halton	Kriging	0	exponential	]250-500]	]0-500]	linear	ML	NelderMead	1.0000000
1190	[5-15]	halton	Kriging	0	gauss	]0-250]	]0-500]	linear	LOO	NelderMead	1.0000000
1270	[5-15]	sobol	Kriging	0	exponential	]0-250]	]0-500]	linear	LOO	BFGS	1.0000000
1368	[5-15]	lhs	Kriging	0	gauss	]250-500]	]0-500]	const	ReML	BFGS	1.0000000
1429	[5-15]	lhs	Kriging	0	matern3/2	]0-250]	]500-1000]	linear	ReML	NelderMead	1.0000000
1438	[5-15]	lhs	Kriging	0	matern7/2	]250-500]	]0-500]	linear	LOO	BFGS	1.0000000
1536	[5-15]	halton	Kriging	0	matern5/2	]250-500]	]500-1000]	const	ML	BOBYQA	1.0000000
1606	[5-15]	halton	Kriging	0	matern5/2	]0-250]	]0-500]	const	ReML	BOBYQA	1.0000000
1624	[5-15]	sobol	Kriging	0	matern3/2	]0-250]	]500-1000]	const	ML	NelderMead	1.0000000
165	[38-50]	sobol	Kriging	0	matern7/2	]250-500]	]1000-1500]	const	ReML	NelderMead	1.0000000
174	[38-50]	sobol	Kriging	0	matern7/2	]0-250]	]0-500]	linear	LOO	NelderMead	1.0000000
262	[38-50]	sobol	Kriging	0	matern5/2	]250-500]	]500-1000]	const	ML	BFGS	1.0000000
927	[38-50]	sobol	Kriging	0	gauss	]250-500]	]500-1000]	const	ReML	BOBYQA	1.0000000
1064	[38-50]	sobol	Kriging	0	gauss	]250-500]	]0-500]	linear	ML	NelderMead	1.0000000
1437	[38-50]	lhs	Kriging	0	matern7/2	]250-500]	]500-1000]	const	LOO	NelderMead	1.0000000
947	[27-37]	halton	Kriging	0	matern5/2	]250-500]	]0-500]	const	ML	BOBYQA	1.0000000
1353	[27-37]	halton	Kriging	0	matern3/2	]250-500]	]0-500]	linear	LOO	BOBYQA	1.0000000
1413	[27-37]	halton	Kriging	0	matern7/2	]250-500]	]1000-1500]	const	ReML	BFGS	1.0000000

Showing 1 to 25 of 1,756 entries

FIGURE 4.17 – IHM : fenêtre de prédiction des configurations performantes

**Prédiction configurations adaptées pour les objectifs définis par l'utilisateur**

**Choix des variables objectives et niveaux à atteindre**

Variables cibles      Modalités souhaitées

LooQ2      "0.90-0.99"     

LooNRMSE      "0.01-0.05"

**Résultats des prédictions de configuration**

	Nb_Exp	DoE_Type	Meta_Typé	PC_Degree	Corr_Func	Scr_Size	Kniter	Trend	Kcrit	Koptalg	Pr
109	[5-15]	halton	Kriging	0	matern7/2	]0-250]	]1000-1500]	linear	ML	BOBYQA	1.0000000
156	[5-15]	sobol	Kriging	0	matern5/2	]250-500]	]0-500]	const	LOO	NelderMead	1.0000000
634	[5-15]	halton	Kriging	0	exponential	]250-500]	]1000-1500]	const	ML	BOBYQA	1.0000000
787	[5-15]	halton	Kriging	0	matern7/2	]250-500]	]500-1000]	linear	ML	NelderMead	1.0000000
794	[5-15]	halton	Kriging	0	matern7/2	]250-500]	]0-500]	linear	LOO	BOBYQA	1.0000000
902	[5-15]	halton	Kriging	0	gauss	]250-500]	]500-1000]	linear	LOO	BOBYQA	1.0000000
1031	[5-15]	halton	Kriging	0	matern3/2	]250-500]	]500-1000]	const	ReML	BFGS	1.0000000
1092	[5-15]	halton	Kriging	0	exponential	]250-500]	]0-500]	linear	ML	NelderMead	1.0000000
1190	[5-15]	halton	Kriging	0	gauss	]0-250]	]0-500]	linear	LOO	NelderMead	1.0000000

Showing 1 to 9 of 1,756 entries

**Sélection et édition de la configuration**

Type de PEN	Nombre d'expérience	Type de métamodèle	Fonction de corrélation	Fonction de tendance	Critère d'optimisation	Algorithme d'optimisation	Nombre d'itérations	taille de l'échantillon de screening
Halton	15	Krigeage	Matern7/2	Linéaire	ML	BOBYQA	1500	250

      Valider la configuration avant exécution

### 4.2.7 Vérification et exécution de la configuration proposée

L'utilisateur sélectionne l'une de ces propositions, et peut la modifier s'il le souhaite. La nouvelle configuration de processus de PEN est alors validée et exécutée. L'utilisateur choisi ici la première configuration proposée, associée à une probabilité d'atteindre les objectifs de 1 et proposant le moins d'expériences nécessaires. Cette configuration est détaillée dans le Tableau 4.9.

TABLEAU 4.9 – Caractéristiques principales de la nouvelle configuration complète du processus de PEN validée par l'utilisateur

Caractéristiques de configuration	Valeur
Objectif	production d'un modèle de substitution
Précision minimale exigée (LooNRMSE)	0.01
Prédictivité minimale exigée (LooQ <sup>2</sup> )	0.98
Stratégie de planification d'expériences	Statique
Modèle numérique	Rotor1a
Type de plan initial	halton
Nombre d'expériences	15
Type de modèle empirique	Krigeage
<b>Paramètres du modèle de krigeage</b>	
Fonction de corrélation	matern7/2
Fonction de tendance	linéaire
Critère d'optimisation	Maximum de vraisemblance
Algorithme d'optimisation	BOBYQA
Nombre d'itérations	1500
Taille de l'échantillon de screening	250

Une fois l'exécution terminée (Figure 4.18), la configuration, ses résultats et ses performances réelles sont capitalisées. Cette capitalisation enrichie la base de connaissances, pour aboutir à une amélioration de l'apprentissage des réseaux bayésiens futurs, utilisés pour les situations à venir.

La configuration a bien une précision LooNRMSE comprise entre 0.01 et 0.05, mais une prédictivité Q<sup>2</sup> supérieure à 0.99. Le résultat pour LooQ<sup>2</sup> avait une probabilité de 0.75 d'appartenir à la modalité "0.90-0.99" réellement et d'être prédit comme appartenant à cette même modalité, compte tenus des indicateurs issus de la matrice de confusion (Table 4.8). En revanche, les cas prédit comme n'étant pas dans "0.90-0.99" avait une probabilité de 0.978 d'être effectivement hors de cette modalité. Cette erreur de prédiction du réseau bayésien est, dans ce contexte, sans gravité. Le modèle de substitution issu de ce processus de PEN est plus prédictif que prévu.

Il est important de préciser que l'utilisateur prend un risque en choisissant, suivant les prédictions du réseau bayésien, une configuration particulière. Il peut être nécessaire,

FIGURE 4.18 – Exécution de la nouvelle configuration issue de l'utilisation du système d'aide à la décision.

```

*****
** TKriging::printLog[]
*****
Input Variables:      epaisseur_patin:longueur_patin:hauteur:longueur_raidisseur:E_elas:z_s
Output Variable:     Freq1
Deterministic trend: linear
Trend parameters (9): [1.779362754984e+01; -2.295324279912e-01; 4.730159986839e-02; -1.3239
2542e+01; -9.865556117009e-02; -1.072704293005e-04; 7.830498729955e-03 ]
Correlation function: URANIE::Modeler::TMatern72CorrFunction
Correlation length:  normalised (not normalised)
                    1.000000e+01 (2.866387e-02 )
                    1.000000e+01 (2.893995e-01 )
                    1.000000e+01 (9.210698e-02 )
                    1.000000e+01 (4.826304e-02 )
                    3.762596e-01 (1.442677e+06 )
                    1.000000e+01 (2.758801e-01 )
                    1.000000e+01 (3.772626e-01 )
                    1.000000e+01 (1.665721e+03 )

Variance of the gaussian process:      0.318606
RMSE (by Leave One Out):                0.302749
Q2:                                      0.997754
*****
NLoorMSE : 0.0106120025868 ; LooQ2 : 0.997754436285
#####
case number 0 : halton DoE type with 15 experiments, Kriging metamodel (CorrFunc matern7/2)
#####Saving in csv#####
line written in csv

```

dans certains contextes industriels, de pondérer la probabilité de mauvaise classification par la gravité de cette erreur pour en déduire un risque. Par exemple, si le système prédit qu'une configuration de PEN particulière aura un  $LooQ^2 > 0.99$  est qu'il en réalité de 0.95, le risque est faible, le modèle de substitution sera tout de même prédictif. En revanche, s'il est en réalité de 0.10, alors les ressources dépensées lors de la réalisation des expériences numériques auront été perdues. L'analyse des matrices de confusions et des mesures d'erreurs est donc essentielle pour appréhender ce risque. Les résultats affichés à l'utilisateur via l'IHM sont accompagnés d'un rapport donnant une interprétation des résultats du système d'aide à la décision (Figure 4.19).

Les concepteurs, par l'utilisation de ce système d'aide à la configuration des processus de PEN, ont pu rapidement établir une configuration de processus de PEN minimisant le nombre de simulations nécessaires, pour produire un modèle de substitution aussi précis et prédictif que désiré, du modèle de comportement dynamique du pulseur d'air. Ce modèle de substitution pourra être réutilisé pour optimiser le comportement du produit et prendre des décisions plus rapidement. De plus, la structure du réseau bayésien donne à l'utilisateur une représentation claire de l'influence des différents éléments d'une configuration sur les résultats attendus. Il pourra alors se concentrer sur les variables de configuration influentes pour améliorer les performances du processus.

Toutes les données relatives au processus de PEN exécuté sont sauvegardées dans la base

FIGURE 4.19 – IHM : fenêtre des résultats du système d’aide à la configuration du processus de PEN

**Résultats de l'exécution de la configuration**

**Bilan**

ID_Produit	ID_Projet	ID_PEN	Auteur	Durée	Nbre Exp.	Type de modèle	Précision	Prédictivité
BE_ROTOR_1_a	Project_X	NDoE_process_1	G.BLONDET	00h30	15	Krigeage	0.0106	0.9978

Rapport d'exécution:  
 La précision du modèle produit a atteint le niveau désiré.  
 La prédictivité n'a pas atteint le niveau désiré. La modalité désirée de LooQ<sup>2</sup> est "0.90-0.99". La modalité réellement observée est ">0.99".  
 La gravité de cette erreur est négative. Le résultat est donc meilleur que prévu.  
 Selon les contraintes définies, le modèle produit peut-être utilisé comme un métamodèle du modèle Rotor 1a, pour la réponse "Freq1".

de connaissances. L'ajout systématique des nouveaux processus de PEN permet d'enrichir la base de connaissances et de progressivement améliorer les prédictions du système d'aide à la décision.

### 4.3 Conclusion de l'application

Le chapitre 4 a mis en application le système proposé pour fournir une aide à la configuration des processus de PEN. Les processus de PEN peuvent nécessiter une expertise avancée pour être maîtrisés. Cette expertise n'étant pas systématiquement présente en entreprise, le système présenté dans ce chapitre a été conçu pour guider l'utilisateur dans ces choix de configuration. Ce système est basé sur une capitalisation structurée des connaissances à l'aide d'une ontologie, et sur une réutilisation de ces connaissances pour prédire des configurations performantes.

Le cas d'étude a illustré le processus d'aide à la décision considéré, par l'utilisation d'un modèle numérique issu de l'industrie automobile. Pour les besoins de cette application, une base de données conséquente a été générée sur les différents modèles numériques disponibles, pour ensuite être structurée par l'ontologie. L'ontologie, en plus des fonctions qu'elle peut offrir en terme de structuration et de contrôle des données, permet de capitaliser des connaissances au sein d'une ou plusieurs entreprises étendues.

Suite à l'extraction des données pertinentes par rapport au problème de l'utilisateur, un réseau bayésien est construit. Une fois le réseau bayésien défini, le système peut prédire d'une part, la performance d'une configuration donnée, et d'autre part, une configuration

probablement performante selon le problème spécifié par l'utilisateur.

Le choix de l'utilisation d'une ontologie et d'un réseau bayésien permet de palier à l'éventuel manque de données :

- le choix de l'ontologie permet d'assurer une capitalisation des connaissances sur tout un ensemble d'entreprises étendues et d'utilisateurs. Les connaissances sont structurées de façon à être partageables et comprises par le plus grand nombre. Cette capacité conduit à capter des connaissances multi-provenances, en plus grande quantité que quand le cas d'une utilisation locale, et avec une grande diversité. Cette approche est adaptée à une entreprise étendue.
- le réseau bayésien permet d'intégrer les connaissances expertes. En cas de manque de données capitalisées, l'introduction de ces connaissances permet d'utiliser le réseau bayésien en premier lieu sans aucune donnée. La capitalisation progressive des données peut ensuite enrichir le système proposé, rendre ses prédictions plus précises et plus fiables, et enfin aboutir à la découverte de nouvelles connaissances.
- les indicateurs de performances prédictives présentés permettent de rendre compte de la qualité et de la fiabilité des prédictions du système. Ces indicateurs permettent aux utilisateurs de conserver un esprit critique sur ces prédictions et de garder le contrôle de la phase de configuration du processus de [PEN](#).

La principale piste d'amélioration concerne la méthode d'apprentissage et les techniques de préparation des données. L'application de différentes méthodes d'apprentissage (apprentissage par renforcement, apprentissage non-supervisés, comme les méthodes de clustering [Fahad 14]) permettront de rendre le système plus robuste et plus automatisable.

L'IHM proposé permet à l'utilisateur de garder le contrôle de tout le processus de décision. Il dispose d'une représentation graphique claire des influences des différentes variables de la configuration d'un processus de [PEN](#) sur ses performances. Il dispose également de mesures de fiabilité des prédictions, et il peut modifier manuellement la configuration proposée avant de l'exécuter.

Ce système d'aide à la configuration permet à la fois de partager des connaissances expertes et de les réutiliser pour réduire le temps et/ou le coût du processus de conception de produits. L'utilisation de ce système rend la mise en place d'un processus de [PEN](#) plus simple et plus rapide. L'utilisateur n'a pas besoin de disposer d'une expertise avancée pour utiliser ces méthodes.





# Conclusion



L'utilisation des outils numériques pour le traitement et l'échange des données et connaissances est un véritable enjeu pour l'industrie mécanique. Dans les secteurs de l'automobile, du ferroviaire, de l'aéronautique, du spatial, du naval, de l'énergie, et de tous les autres secteurs impactés par la mécanique, la simulation numérique, et les outils de gestion des données et processus liés aux produits, occupent une place de plus en plus importante. Les activités de conception trouvent en ces outils les moyens pour aboutir à un produit répondant à de multiples exigences fortes, tout en étant plus robuste, plus fiable, innovant, optimisé, développé rapidement et à moindre coût, dans un contexte fortement collaboratif et d'entreprise étendue. Réaliser de meilleurs produits, plus durables, plus performants, avec moins d'énergie est un but commun de tous les acteurs de l'industrie mécanique.

L'utilisation de plus en plus fréquente des outils de simulation, de modèles de plus en plus complexes et précis, mobilisant de multiples expertises différentes au cours du développement du produit, mettent les entreprises face à un besoin de gestion des données et des connaissances spécifiques pour la simulation numérique et pour la gestion du cycle de vie du produit virtuel.

La prise en compte des variabilités et des incertitudes du produit directement dans les modèles de simulation est réalisée notamment par l'utilisation de plans d'expériences numériques (PEN). Une série de simulations est planifiée pour simuler la variation des propriétés, de l'environnement et des performances du produit pour l'améliorer plus tôt dans le processus de conception et se passer autant que possible de prototypes. L'utilisation des PEN nécessite pourtant la mise en place d'une démarche complexe. Les concepteurs doivent établir une stratégie spécifique à leur produit pour définir en avance la meilleure manière d'appliquer une démarche de PEN pour obtenir les résultats désirés tout en maîtrisant les coûts, les temps de simulation et la quantité de données générées. Même si les moyens de calculs ont largement évolués ces dernières décennies, avec l'arrivée des plateformes de calcul HPC et des calculateurs exaflopiques au sein des entreprises, cela ne compense pas les besoins pour traiter des produits complexes soumis à de fortes exigences.

Des stratégies et des méthodes existent pour appliquer efficacement un processus de PEN à des modèles numériques complexes. Par exemple, une stratégie adaptative permet de minimiser le nombre de simulations pour obtenir les résultats désirés. Toutefois, ces stratégies améliorent d'autant plus l'efficacité des PEN qu'elles requièrent des connaissances expertes plus avancées. Les processus de PEN deviennent alors plus complexes à mettre en place pour des simulations complexes et coûteuses.

Ces travaux, présentés dans ce manuscrit, se sont attachés à proposer une solution pour faciliter l'application des processus de PEN au sein de démarches de conception de produits. L'objectif est de pouvoir obtenir rapidement la définition la plus efficace d'un

processus de **PEN** pour un problème donné et pour atteindre les objectifs de l'étude.

Certaines solutions ont été proposées en ce sens, mais celle envisagée dans ces travaux se veut être plus exhaustive, plus réactive et apporte une meilleure gestion des connaissances capitalisées en entreprises. Le premier axe de recherche s'est porté vers la capitalisation et la structuration des connaissances propres aux **PEN**. Un manque a été identifié parmi tous les développements existants en termes de modélisation et de structuration des connaissances, au croisement entre les domaines de la statistique et de l'ingénierie mécanique. De ces travaux, une modélisation ontologique a été produite, pour assurer la capitalisation et le partage de l'expertise propre au **PEN** dans un contexte de développement de produit. Le second axe a traité de la réutilisation des connaissances, et notamment les systèmes à base de connaissances. De nombreuses méthodes, non-appliquées aux processus de **PEN**, existent dans le domaine de l'Intelligence Artificielle. Depuis les premiers systèmes experts développés au début des années 1990, des systèmes plus évolués, capables d'apprendre par eux-mêmes ont émergé. Parmi les méthodes couvertes dans l'état de l'art, les réseaux bayésiens ont été identifiés comme étant les plus adaptés pour répondre à notre problématique.

Le système à base de connaissances proposé repose sur deux éléments. Le premier élément est une ontologie, ODE (Ontology for Design of Experiments). ODE apporte une structuration des connaissances spécifiques aux **PEN**, jusqu'alors inexistantes, produite en cohérence avec les développements existants pour la gestion des connaissances liées au développement de produits. ODE offre une base pour la capitalisation, le partage, la diffusion et la réutilisation de connaissances pour l'application de démarches de **PEN** au sein du développement de produits dans un contexte de conception collaborative et multi-disciplinaire.

Le second élément est un moteur d'inférences basé sur les réseaux bayésiens. Le but du moteur d'inférences est de fournir des prédictions de l'efficacité d'une configuration de processus de **PEN** donnée, ou de prédire la configuration la plus efficace pour un contexte donné. Ce type de moteur d'inférences est capable de supporter les connaissances déjà existantes dans l'entreprise, d'apprendre à partir des résultats des précédents processus de **PEN**, de gérer tous les éléments caractérisant un processus de **PEN** (sa configuration et ses résultats), de fournir des raisonnements explicites et de proposer des solutions inédites pour la configuration du processus de **PEN**. Sa capacité à combiner les connaissances expertes, si elles existent, et l'apprentissage automatique lui permet d'actualiser les connaissances capitalisées, de les compléter en exploitant les données des processus historisés, et d'explorer de nouvelles solutions. De plus, le réseau bayésien est muni d'indicateurs de performances prédictives, quantifiant sa capacité à réaliser des prédictions correctes. L'utilisateur peut alors évaluer les chances de réussite d'une configuration de processus de **PEN** avant son exécution.

Cette combinaison originale forme un système à base de connaissances permettant d'intégrer les connaissances de l'entreprise de manière structurée et compréhensible, de les communiquer et de les exploiter pour accélérer les démarches de conception de produit. Ce système à base de connaissances offre une assistance aux concepteurs pour la mise en place de processus de **PEN** efficaces, en réutilisant les connaissances de l'entreprise, mais aussi en explorant d'autres solutions innovantes jamais essayées auparavant. Ce système propose des processus de **PEN** entièrement configurés adaptés aux besoins et aux ressources du projet. Sa capacité à utiliser à la fois les connaissances et les données disponibles lui permet d'être utilisé à la fois pour mettre à profit les savoir-faire de l'entreprise et pour proposer de nouvelles solutions innovantes et plus efficaces. Ce système est également capable d'expliquer ses raisonnements et de mettre à disposition des concepteurs de nouvelles connaissances compréhensibles. Le cas d'application sur un produit issu du secteur automobile a montré sa capacité à s'adapter et à proposer de nouvelles configurations de **PEN** efficaces et économiques dans un contexte spécifique. Il fournit lui-même des estimations de ces capacités à assister les concepteurs. Cette capacité place les concepteurs au centre des décisions, en leurs donnant des indicateurs d'estimations des risques pris par l'utilisation de démarches de **PEN** novatrices. Enfin, ce système s'enrichit de l'intégralité des succès et des échecs d'applications des **PEN** et s'améliore constamment. Cette dernière capacité est elle-même améliorée par ODE, qui favorise l'échange des connaissances et des données dans des contextes collaboratifs.

Le système proposé accélère la mise en place et améliore l'efficacité de méthodes statistiques avancées pour un développement rapide de produits complexes, en réutilisant et en créant des connaissances au sein de l'entreprise étendue. Les concepteurs auront alors une meilleure maîtrise et une meilleure compréhension du processus de **PEN**. Ils profiteront donc de ces méthodes pour améliorer leur compréhension du comportement du produit et l'optimiser.

Ce système à base de connaissances peut toutefois être encore amélioré suivant plusieurs axes.

La premier axe concerne l'amélioration de la capacité d'apprentissage du moteur d'inférences. De nombreux aspects de l'IA et des réseaux bayésiens restent à exploiter. Les réseaux bayésiens peuvent par exemple gérer des connaissances et données incomplètes et manquantes par des algorithmes spécifiques. De nombreuses méthodes d'apprentissages des paramètres et de la structure des réseaux bayésiens existent et pourront être testées. Par exemple, un mode d'apprentissage par renforcement pourrait être ajouté afin de donner plus d'importance aux processus de **PEN** rarement utilisés mais très efficaces ou totalement inefficaces. Cela améliorerait la capacité d'exploration de nouvelles solutions et ses performances prédictives. De plus, des méthodes d'apprentissage non-supervisés ou d'autres méthodes de classification pourraient faciliter le choix des variables et de leurs modalités,

rendant l'application du réseau bayésien plus rapide. Enfin, de nombreux algorithmes d'apprentissage de la structure du réseau existent et pourraient être testés.

Le second axe concerne l'extension des réseaux bayésiens à d'autres méthodes. Les capacités prédictives des réseaux bayésiens devraient être comparées avec celles des arbres de décisions, des forêts aléatoires et des réseaux de neurones sur des cas d'applications précis. Ces systèmes pourraient être combinés et former un système hybride, à l'image des nombreux développements de ces dernières années dans ce domaine. De plus, l'aspect collaboratif pourrait être renforcé en créant des interactions entre plusieurs instances du système à base de connaissances, chaque instance étant liée à un utilisateur ou à une entité de l'entreprise étendue. Cela permettrait de confronter les prédictions de plusieurs réseaux bayésiens basés sur des bases de connaissances différentes et d'engendrer une décision finale, basée sur, par exemple, un vote à la majorité entre tous les réseaux bayésiens.

Le troisième axe est la validation de cette proposition dans un contexte de travail collaboratif industriel pour la conception de produits complexes. L'échange des connaissances entre entités et organisations différentes, concurrentes ou soumises à des exigences de sécurité est possible concernant l'application de processus de [PEN](#). L'anonymat des connaissances peut être assuré, ouvrant des possibilités de conception d'une base de connaissances unifiée et mondialisée. Cet axe est actuellement en voie d'application dans le groupe Valeo (groupe présent dans 32 pays, comprenant notamment 155 sites de productions et 58 centres de recherches et de développements), faisant suite au projet SDM4DOE, pour une utilisation des [PEN](#) intensifiée au sein des équipes de développement, en s'appuyant sur le système à base de connaissances proposé ici.

# Bibliographie

- [Aamodt 94] Agnar Aamodt & Enric Plaza. *Case-based reasoning : Foundational issues, methodological variations, and system approaches*. AI communications, vol. 7, no. 1, pages 39–59, 1994. [83](#)
- [Abburu 12] Sunitha Abburu. *A Survey on Ontology Reasoners and Comparison*. International Journal of Computer Applications, vol. 57, no. 17, pages 33–39, 2012. [68](#)
- [Adams 08] Vince Adams. *A Designer’s Guide to Simulation with Finite Element Analysis*. NAFEMS, 2008. [16](#)
- [AFM 15] AFM. *Livre blanc de la recherche en mécanique*. Association Française de Mécanique, EDP Sciences, 2015. [8](#), [18](#)
- [Allaire 05] G. Allaire. *Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique*. Mathématiques appliquées. Editions de l’Ecole polytechnique, 2005. [40](#)
- [Arlat 96] J. Arlat, J.P. Blanquart, A. Costes, A. Crouzet, Y. Deswarte, J.C. Fabre, H. Guillermain, M. Kaâniche, K. Kanoun, J.C. Laprie, C. Mazet, D. Powell, C. Rabéjac & J. Thévenoud. *Guide de la Sûreté de Fonctionnement*. 2nd edition, 1996. [18](#)
- [Assouroko 14] Ibrahim Assouroko, Guillaume Ducellier, Philippe Boutinaud & Benoît Eynard. *Knowledge management and reuse in collaborative product development – a semantic relationship management-based approach*. International Journal of Product Lifecycle Management, vol. 7, no. 1, pages 54–74, 2014. [70](#)
- [Baillargeon 05] Sophie Baillargeon. *Le krigeage : revue de la théorie et application à l’interpolation spatiale de données de précipitations*. PhD thesis, Université Laval, Québec, 2005. [36](#)
- [Barbau 12] Raphael Barbau, Sylvere Krима, Rachuri Sudarsan, Anantha Narayanan, Xenia Fiorentini, Sebti Foufou & Ram D. Sriram. *OntoSTEP : Enriching product model data using ontologies*. Computer-Aided Design, vol. 44, no. 6, pages 575–590, 2012. [70](#)
- [Beal 14] A. Beal, Magalie Claeys-Bruno & M. Sergent. *Constructing space-filling designs using an adaptive WSP algorithm for spaces with constraints*. Chemometrics and Intelligent Laboratory Systems, vol. 133, pages 84–91, apr 2014. [27](#)
- [Benhafid 05] Yassine Benhafid, Nadège Troussier, Nassim Boudaoud & Zohra Cherfi. *Méthode d’aide à l’idéalisaton de modèles issus de la CAO pour le calcul de structures*. Mechanics & Industry, vol. 6, no. 3, pages 289–295, may 2005. [17](#)

- [Boussaïd 13] Ilhem Boussaïd, Julien Lepagnot & Patrick Siarry. *A survey on optimization metaheuristics*. Information Sciences, vol. 237, pages 82–117, jul 2013. [40](#)
- [Box 59] G E P Box & Norman R Draper. *A Basis for the Selection of a Response Surface Design*. Journal of the American Statistical Association, vol. 54, no. 287, pages 622–654, 1959. [27](#)
- [BPI 14] BPI. *Cœur de filière numérique - Calcul intensif et simulation numérique n°2*, 2014. [11](#)
- [Brangé 14] Jean Brangé, Everson Moura & Jean-Yves Delaunay. *Development of STEP AP 242 ed2 “Managed Model Based 3D Engineering”*, 2014. [60](#)
- [Breiman 01] Leo Breiman. *Random Forests*. Machine Learning, vol. 45, no. 1, pages 5–32, 2001. [80](#)
- [Bruno 15] Giulia Bruno, Dario Antonelli & Agostino Villa. *A reference ontology to support product lifecycle management*. In 9th CIRP Conference on Intelligent Computation in Manufacturing Engineering - CIRP ICME '14, volume 33, pages 41–46. Elsevier B.V., 2015. [15](#)
- [Cai 11] Zhiqiang Cai, Shudong Sun, Shubin Si & Bernard Yannou. *Identifying product failure rate based on a conditional Bayesian network classifier*. Expert Systems with Applications, vol. 38, no. 5, pages 5036–5043, may 2011. [82](#)
- [Cameron 47] Robert H Cameron & William T Martin. *The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals*. Annals of Mathematics, pages 385–392, 1947. [34](#)
- [CETIM 15] CETIM. *Technologies prioritaires 2020 en Mécanique*, 2015. [10](#)
- [Chandrasegaran 13] Senthil K. Chandrasegaran, Karthik Ramani, Ram D. Sriram, Imre Horváth, Alain Bernard, Ramy F. Harik & Wei Gao. *The evolution, challenges, and future of knowledge representation in product design systems*. Computer-Aided Design, vol. 45, no. 2, pages 204–228, feb 2013. [53](#), [54](#)
- [Chastaing 13] Gaelle Chastaing. *Indices de Sobol généralisés pour variables dépendantes*. PhD thesis, Université de Grenoble, 2013. [37](#)
- [Chen 91] Frederick T. Chen. *A personal computer based expert system framework for the design of experiments*. Computers & Industrial Engineering, vol. 21, pages 197–200, 1991. [78](#)
- [Chen 06] Victoria C.P. Chen, Kwok-Leung Tsui, Russell R. Barton & Martin Meckesheimer. *A review on design, modeling and applications of computer experiments*. IIE Transactions, vol. 38, no. 4, pages 273–291, apr 2006. [30](#)
- [Cheung 10] Yan-Leung Cheung, Ping Jiang, Piman Limpaphayom & Tong Lu. *Corporate Governance in China : a Step Forward*. European Financial Management, vol. 16, no. 1, pages 94–123, jan 2010. [53](#)
- [CIMDATA 03] CIMDATA. *Product Lifecycle Management —Empowering the future of business*. Rapport technique, CIMDATA, 2003. [13](#), [14](#)
- [CIMDATA 11] CIMDATA. *Simulation Lifecycle Management*. Rapport technique September, CIMDATA, 2011. [43](#)



- [CIMDATA 12] CIMDATA. *Beyond Simulation Data Management*. Rapport technique September, CIMDATA, 2012. [13](#), [43](#), [62](#)
- [CIMDATA 14] CIMDATA. *Simulation & Analysis Governance*. Rapport technique April, CIMDATA, 2014. [62](#)
- [CIMDATA 16] CIMDATA. *PLM Road Map™ 2016 for The Automotive Industry*. Rapport technique, CIMDATA, 2016. [14](#)
- [Correa 09] M. Correa, C. Bielza & J. Pamies-Teixeira. *Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process*. Expert Systems with Applications, vol. 36, no. 3, pages 7270–7279, 2009. [82](#)
- [CRESCENDO 13] CRESCENDO. *CRESCENDO Forum Handbook*. In CRESCENDO Forum, Toulouse, 2013. [15](#), [43](#)
- [Crestaux 09] Thierry Crestaux, Olivier Le Maître & Jean-Marc Martinez. *Polynomial chaos expansion for sensitivity analysis*. Reliability Engineering & System Safety, vol. 94, no. 7, pages 1161–1172, jul 2009. [34](#), [35](#)
- [Crombecq 11] Karel Crombecq, Dirk Gorissen, Dirk Deschrijver & Tom Dhaene. *A Novel Hybrid Sequential Design Strategy for Global Surrogate Modeling of Computer Experiments*. SIAM Journal on Scientific Computing, vol. 33, no. 4, pages 1948–1974, 2011. [74](#)
- [Cui 16a] Can Cui, Mengqi Hu, Jeffery D. Weir & Teresa Wu. *A recommendation system for meta-modeling : A meta-learning based approach*. Expert Systems with Applications, vol. 46, pages 33–44, mar 2016. [32](#)
- [Cui 16b] Can Cui, Mengqi Hu, Jeffery D. Weir & Teresa Wu. *A recommendation system for meta-modeling : A meta-learning based approach*. Expert Systems with Applications, vol. 46, pages 33–44, mar 2016. [79](#)
- [Dalkir 05] Kimiz Dalkir. *Knowledge Management in Theory and Practice*. Elsevier B.V., 2005. [53](#), [54](#)
- [Danjou 14] Christophe Danjou, Julien Le Duigou & Benoît Eynard. *OntoSTEP-NC for Information Feedbacks from CNC to CAD/CAM Systems*. Advances in Production Management Systems., vol. 440, pages 256–263, 2014. [70](#)
- [de Rocquigny 08] Etienne de Rocquigny, Nicolas Devictor & Stefano Tarantola. *Uncertainty in industrial practice : a guide to quantitative uncertainty management*. John Wiley & Sons, 2008. [38](#)
- [Denis 14] Jean-Baptiste Denis & Marco Scutari. *Réseaux bayésiens avec R*. EDP Sciences, 2014. [114](#), [117](#), [133](#), [150](#)
- [Ding 09] Lian Ding & Jason Matthews. *A contemporary study into the application of neural network techniques employed to automate CAD/CAM integration for die manufacture*. Computers and Industrial Engineering, vol. 57, no. 4, pages 1457–1471, 2009. [79](#)
- [Dolšak 02] B Dolšak. *Finite element mesh design expert system*. Knowledge-Based Systems, vol. 15, no. 5-6, pages 315–322, 2002. [78](#)
- [Draghici 09] George Draghici & Anca Draghici. *Collaborative Product Development in PLM Multisite Platform*. In 1st International Conference on Manufacturing Engineering, Quality and Production Systems (Volume II), volume II, pages 327–332, 2009. [15](#)

- [Dreyfus 11] Gérard Dreyfus, Jean-Marc Martinez, Manuel Samuelides, M. B. Gordon, F. Badran & S. Thiria. Apprentissage statistique : Réseaux de neurones-Cartes topologiques-Machines à vecteurs supports. Eyrolles, 2011. 31, 35, 146
- [Dubourg 11] Vincent Dubourg. *Méta-modèles adaptatifs pour l'analyse de fiabilité et l'optimisation sous contrainte fiabiliste*. PhD thesis, Université Blaise Pascal-Clermont II, 2011. 28
- [Fahad 14] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou & Abdelaziz Bouras. *A Survey of Clustering Algorithms for Big Data : Taxonomy and Empirical Analysis*. IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pages 267–279, sep 2014. 14, 117, 163
- [Faure 82] Henri Faure. *Discrépance de suites associées à un système de numération (en dimension  $s$ )*. Acta Arithmetica, vol. 41, no. 4, pages 337–351, 1982. 27
- [Feilmayr 16] Christina Feilmayr & Wolfram Wöb. *An analysis of ontologies and their success factors for application to business*. Data & Knowledge Engineering, vol. 101, pages 1–23, jan 2016. 65
- [Fenves 03a] Steven J Fenves, Ram D. Sriram, Young Choi, Balan Gurumoorthy & Greg Mocko. *Master Product Model for the Support of Tighter Integration of Spatial and Functional Design*. Rapport technique, National Institute of Standards and Technology (NIST), 2003. 59
- [Fenves 03b] Steven J Fenves, R.D. Sriram, Rachuri Sudarsan & F. Wang. *A product information modeling framework for product lifecycle management*. In International Symposium on Product Lifecycle Management, volume 37, pages 1399–1411, Bangalore, India, nov 2003. 59
- [Fenves 06] Steven J Fenves, Sebti Foufou, Conrad Bock & Ram D. Sriram. *CPM : A Core Model for Product Data*, 2006. 58
- [Fiorentini 07] Xenia Fiorentini, Iacopo Gambino, Vei-Chung Liang, Rachuri Sudarsan, Mahesh Mani & Conrad Bock. *An Ontology for Assembly Representation*. In NISTIR 7436, page 76, 2007. 59, 70
- [Fowlkes 95] William Y Fowlkes & Clyde M Creveling. Engineering methods for robust product design. Addison-Wesley, 1995. 24
- [Franco 08] Jessica Franco. *Planification d'expériences numériques en phase exploratoire pour la simulation des phénomènes complexes*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne (ENSM-SE), 2008. 27, 28, 30
- [Gazut 08] Stéphane Gazut, Jean-Marc Martinez, Gérard Dreyfus & Yacine Oussar. *Towards the optimal design of numerical experiments*. IEEE transactions on neural networks, vol. 19, no. 5, pages 874–82, may 2008. 26
- [Gorissen 09] Dirk Gorissen, Tom Dhaene & Filip De Turck. *Evolutionary Model Type Selection for Global Surrogate Modeling*. Journal of Machine Learning Research, vol. 10, pages 2039–2078, 2009. 74
- [Goupy 00] Jacques Goupy. *Modélisation par les plans d'expériences*. Techniques de l'ingénieur, 2000. 26

- [Grieves 14] Michael Grieves. *Digital Twin : Manufacturing Excellence through Virtual Factory Replication*. Rapport technique, 2014. [14](#)
- [Grosse 05] Ian R. Grosse, John M. Milton-Benoit & Jack C. Wileden. *Ontologies for supporting engineering analysis models*. AI EDAM, vol. 19, no. 01, pages 1–18, feb 2005. [70](#)
- [Gruber 93a] Thomas R. Gruber. *Toward principles for the design of ontologies used for knowledge sharing*. International Journal of Human-Computer Studies, vol. 43, no. 5-6, pages 907–928, nov 1993. [65](#)
- [Gruber 93b] Thomas R. Gruber. *Toward principles for the design of ontologies used for knowledge sharing*. Formal Ontology in Conceptual Analysis and Knowledge Representation, vol. 43, no. 5-6, pages 907–928, 1993. [100](#)
- [Gruhier 15] Elise Gruhier, Frédéric Demoly, Olivier Dutartre, Said Abboudi & Samuel Gomes. *A formal ontology-based spatiotemporal mereotopology for integrated product design and assembly sequence planning*. Advanced Engineering Informatics, vol. 29, no. 3, pages 495–512, aug 2015. [68](#)
- [Haas 16] K. Haas, H. Schuck, T. Mücke & J. Ovtcharova. *A Holistic Product Lifecycle Management Approach to Support Design by Machine Data*. 26th CIRP Design Conference, vol. 50, pages 420–423, 2016. [15](#)
- [Halton 60] J. H. Halton. *On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals*. Numerische Mathematik, vol. 2, no. 1, pages 84–90, dec 1960. [27](#)
- [Hammersley 60] J. M. Hammersley. *Monte Carlo methods for solving multivariable problems*. Annals of the New York Academy of Sciences, vol. 86, no. 3, pages 844–874, 1960. [27](#)
- [Hanafy 14] Mohmmad Hanafy & Hoda ElMaraghy. *Co-design of Products and Systems Using a Bayesian Network*. 47th CIRP Conference on Manufacturing Systems, vol. 17, pages 284–289, 2014. [82](#)
- [Hastie 09] Trevor Hastie, Robert Tibshirani & Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2 edition, 2009. [32](#), [80](#), [119](#)
- [Hedayat 78] A. Hedayat & W.D. Wallis. *Hadamard matrices and their applications*. The Annals of Statistics, vol. 6, no. 6, pages 1184–1238, 1978. [26](#)
- [Hu 08] Wang Hu, Li Guang Yao & Zhong Zhi Hua. *Optimization of sheet metal forming processes by adaptive response surface based on intelligent sampling method*. Journal of Materials Processing Technology, vol. 197, no. 1, pages 77–88, feb 2008. [28](#)
- [IATA 11] IATA. *Vision 2050*. Rapport technique February, International Air Transport Association, 2011. [8](#)
- [IFIP/IFAC 99] IFIP/IFAC. *GERAM : Generalised Enterprise Reference Architecture and Methodology*. Rapport technique March, IFIP/IFAC, 1999. [12](#)
- [Iooss 14] Bertrand Iooss & Paul Lemaître. *A review on global sensitivity analysis methods*. In Gabriella Dellino & Carlo Meloni, editeurs, Uncertainty Management in Simulation-Optimization of Complex Systems-Algorithms and Applications-Part 2, pages 101–122. Springer US, 2014. [24](#), [32](#), [37](#), [38](#)
- [IRTSystemX 16] IRTSystemX. *L’IRT SystemX dévoile sa feuille de route pour la période 2016-2020 axée autour du développement de projets orientés marchés et usages*. Communiqué de presse, 2016. [11](#)

- [ISO 10] ISO. *ISO 10303-209 Edition 2 – Multidisciplinary Design and Analysis Background and Capabilities High Level Overview and Background*, 2010. 61
- [ISO 14] ISO. *ISO 10303-242 - Application protocol : managed model-based 3D engineering*, 2014. 60
- [ISO 15] ISO. *ISO 10303 (STEP) AP 239 edition 3 Application Protocol For Product Life Cycle Support (PLCS)*, 2015. 59
- [Iuliano 13] Emiliano Iuliano & Domenico Quagliarella. *Proper Orthogonal Decomposition, surrogate modelling and evolutionary optimization in aerodynamic design*. *Computers and Fluids*, vol. 84, pages 327–350, sep 2013. 28
- [Jacques 05] Julien Jacques. *Contributions à l'analyse de sensibilité et l'analyse discriminante généralisée*. PhD thesis, Université Joseph Fourier-Grenoble 1, 2005. 37
- [Jang 93] J.-S.R. Jang. *ANFIS : adaptive-network-based fuzzy inference system*. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pages 665–685, 1993. 84
- [Johnson Mark 90] E Johnson Mark, M Moore Leslie & Donald Ylvisaker. *Minimax and maximin distance designs*. *Journal of statistical planning and inference*, vol. 2, pages 131–148, 1990. 27
- [Jun 07] Hong-bae Jun, Dimitris Kiritsis & Paul Xirouchakis. *Research issues on closed-loop PLM*. *Computers in Industry*, vol. 58, no. 8-9, pages 855–868, dec 2007. 14
- [Kass 80] G. V. Kass. *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 29, no. 2, pages 119–127, 1980. 80
- [Kebede 10] Gashaw Kebede. *Knowledge management : An information science perspective*. *International Journal of Information Management*, vol. 30, no. 5, pages 416–424, 2010. 53
- [Kestel 16] P Kestel, T Schneyer & S Wartzack. *Feature-Based Approach for the Automated Setup of Accurate , Design Accompanying Finite Element Analyses*. In *International design conference - DESIGN 2016*, pages 697–706, Dubrovnik - Croatia, 2016. 74
- [Khan 14] Abid Ali Khan, Imran Ali Chaudhry & Ali Sarosh. *Case Based Reasoning Support for Adaptive Finite Element Analysis : Mesh Selection for an Integrated System*. *Applied Physics Research*, vol. 6, no. 3, pages 21–39, apr 2014. 83
- [Khuri 10] I Khuri & Siuli Mukhopadhyay. *Response surface methodology*. *WIREs Computational Statistics*, vol. 2, 2010. 26
- [Kleijnen 12] Jack P.C. Kleijnen, Wim Van Beers & Inneke Van Nieuwenhuyse. *Expected improvement in efficient global optimization through bootstrapped kriging*. *Journal of Global Optimization*, vol. 54, no. 1, pages 59–73, 2012. 28, 30
- [Le Duigou 10] Julien Le Duigou. *Cadre de modélisation pour les systèmes PLM en entreprise étendue. Application aux PME mécaniciennes*. PhD thesis, Ecole Centrale de Nantes (ECN), 2010. 9, 12, 13

- [Le Duigou 11] Julien Le Duigou & Alain Bernard. *Product Lifecycle Management Model for Design Information Management in Mechanical Field*. In 21st CIRP Design Conference, pages 207–213, Korea, 2011. [70](#), [102](#)
- [Li 03] S Li. *A hybrid expert system for finite element modeling of fuselage frames*. Expert Systems with Applications, vol. 24, no. 1, pages 87–93, 2003. [84](#)
- [Liu 08] Jia-li Liu, Xiang-bin Yan, Wei Qi & Yi-jun Li. *A case-based reasoning system for mechanical design*. 2008 International Conference on Management Science and Engineering 15th Annual Conference Proceedings, pages 585–590, 2008. [83](#)
- [Lopez de Mantaras 05] Ramon Lopez de Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, Michael T. Cox, Kenneth Forbus, Mark Keane, Agnar Aamodt & Ian Watson. *Retrieval, reuse, revision and retention in case-based reasoning*. The Knowledge Engineering Review, vol. 20, no. 03, page 215, sep 2005. [83](#)
- [Lorenzen 92] Thomas J. Lorenzen, Lynn T. Truss, W. Scott Spangler, William T. Corpus & Andrew B. Parker. *DEXPERT : an expert system for the design of experiments*. Statistics and Computing, vol. 2, no. 2, pages 47–54, 1992. [77](#)
- [Lu 15] Wenlong Lu, Yuchu Qin, Xiaojun Liu, Meifa Huang, Liping Zhou & Xiangqian Jiang. *Enriching the semantics of variational geometric constraint data with ontology*. CAD Computer Aided Design, vol. 63, pages 72–85, 2015. [69](#)
- [Maria 97] Anu Maria. *Introduction to modeling and simulation*. In Winter Simulation Conference, pages 7–13, 1997. [16](#)
- [Matheron 67] G Matheron. *Kriging or polynomial interpolation procedures*. CIMM Transactions, vol. 70, pages 240–244, 1967. [36](#)
- [Matsokis 10] Aristeidis Matsokis & Dimitris Kiritsis. *An ontology-based approach for Product Lifecycle Management*. Computers in Industry, vol. 61, no. 8, pages 787–797, oct 2010. [69](#)
- [McKay 79] M. D. McKay, R. J. Beckman & W. J. Conover. *Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code*. Technometrics, vol. 21, no. 2, pages 239–245, 1979. [27](#)
- [Montgomery 00] Douglas C Montgomery. *Design and analysis of experiments*. John Wiley & Sons, Inc., 5th edition, 2000. [26](#)
- [Naïm 07] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret & Anna Becker. *Réseaux bayésiens*. Eyrolles, 3e édition edition, 2007. [79](#), [81](#), [82](#), [113](#), [114](#), [117](#)
- [Naranje 14] V. Naranje & S. Kumar. *A knowledge based system for automated design of deep drawing die for axisymmetric parts*. Expert Systems with Applications, vol. 41, no. 4, pages 1419–1431, mar 2014. [78](#)
- [NASA 15] NASA. *NASA Technology Roadmaps*. Rapport technique, National Aeronautics and Space Administration, 2015. [17](#)

- [Niederreiter 88] Harald Niederreiter. *Low-discrepancy and low-dispersion sequences*. Journal of number theory, vol. 30, no. 1, pages 51–70, 1988. [27](#)
- [Nonaka 01] Ikujiro Nonaka, Ryoko Toyama & Philippe Byosière. *A Theory of Organizational Knowledge Creation : Understanding the Dynamic Process of Creating Knowledge*. In Handbook of Organizational Learning and Knowledge, pages 491–516. Elsevier, New York, NY, 2001. [53](#), [54](#), [55](#)
- [Norris 12] Mark Norris. *Business Value from Simulation Data Management-a Decade of Production Experience*. Rapport technique, NAFEMS, 2012. [45](#)
- [Owen 02] Roderick Owen & Imre Horváth. *Towards Product-Related Knowledge Asset Warehousing in Enterprises*. In Proceedings of the Fourth International Symposium on Tools and Methods of Competitive Engineering, pages 155–170, 2002. [53](#)
- [Panetto 12] H. Panetto, M. Dassisti & A. Tursi. *ONTO-PDM : Product-driven Ontology for Product Data Management interoperability within manufacturing process environment*. Advanced Engineering Informatics, vol. 26, no. 2, pages 334–348, apr 2012. [69](#)
- [Pardessus 00] Thierry Pardessus. *The Multi-Site Extended Enterprise Concept in the Aeronautical Industry*. Aircraft Technologies, 2000. [15](#)
- [Patterson 16] Eann A. Patterson, Richard J. Taylor & Mark Bankhead. *A framework for an integrated nuclear digital environment*. Progress in Nuclear Energy, vol. 87, pages 97–103, mar 2016. [14](#)
- [Pawlak 94] Zdzisaw Pawlak & Roman Sowinski. *Rough set approach to multi-attribute decision analysis*. European Journal of Operational Research, vol. 72, no. 3, pages 443–459, 1994. [84](#)
- [Pearl 00] Judea Pearl. *Causality : Models, Reasoning and Inference*. Cambridge University Press, Cambridge, England, 2000. [114](#)
- [Pease 02] Adam Pease, Ian Niles & John Li. *The Suggested Upper Merged Ontology : A Large Ontology for the Semantic Web and its Applications*. Rapport technique, Association for the Advancement of Artificial Intelligence, 2002. [71](#)
- [Philip Chen 14] C. L. Philip Chen & Chun Yang Zhang. *Data-intensive applications, challenges, techniques and technologies : A survey on Big Data*. Information Sciences, vol. 275, pages 314–347, aug 2014. [31](#)
- [Pourroy 99] Franck Pourroy. *Apport des éléments finis à la conception mécanique*. In Méthodes et outils pour la conception, chapitre AG2530, pages 475–483. Les Techniques de L’Ingénieur, Paris, 1999. [16](#)
- [Powers 11] David M. W. Powers. *Evaluation : from precision, recall and F-measure to ROC, informedness, markedness and correlation*. Journal of Machine Learning Technologies, vol. 2, no. 1, pages 37–63, 2011. [119](#), [121](#)
- [Pratt 05] Michael J. Pratt. *ISO 10303, the STEP standard for product data exchange, and its PLM capabilities*. International Journal of Product Lifecycle Management, vol. 1, no. 1, page 86, 2005. [59](#)
- [Quinlan 96] J. Ross Quinlan. *Improved use of continuous attributes in C4. 5*. Journal of artificial intelligence research, vol. 4, pages 77–90, 1996. [80](#)

- [Rakotomalala 05] Ricco Rakotomalala. *Arbres de Décision*. La Revue de Modulad (33), pages 163–187, 2005. [80](#)
- [Rehage 15] Gerald Rehage & Jürgen Gausemeier. *Ontology-based Determination of Alternative CNC Machines for a Flexible Resource Allocation*. In 15th CIRP Conference on Modelling of Machining Operations, volume 31, pages 47–52. Elsevier B.V., 2015. [70](#)
- [Robin 07] Vincent Robin, Bertrand Rose & Philippe Girard. *Modelling collaborative knowledge to support engineering design project manager*. Computers in Industry, vol. 58, no. 2, pages 188–198, 2007. [53](#)
- [Robinson 77] R W Robinson. *Counting unlabeled acyclic digraphs*. Lecture Notes in Mathematics, vol. 622, pages 28–43, 1977. [114](#)
- [Roy 13] Shibendu Shekhar Roy & Dilip Kumar Pratihar. *Adaptive neuro-fuzzy expert systems for predicting specific energy consumption and energy stability margin in crab walking of six-legged robots*. Journal of Intelligent and Fuzzy Systems, vol. 24, no. 3, pages 467–482, 2013. [84](#)
- [Russel 10] Stuart Russel & Peter Norvig. *Artificial Intelligence - A modern approach*. 3rd edition, 2010. [75](#), [114](#)
- [Sahin 12] S. Sahin, M.R. Tolun & R. Hassanpour. *Hybrid expert systems : A survey of current approaches and applications*. Expert Systems with Applications, vol. 39, no. 4, pages 4609–4617, mar 2012. [76](#), [84](#)
- [Sakthivel 10] N. R. Sakthivel, V. Sugumaran & S. Babudevasenapati. *Vibration based fault diagnosis of monoblock centrifugal pump using decision tree*. Expert Systems with Applications, vol. 37, no. 6, pages 4040–4049, 2010. [80](#), [81](#)
- [Saltelli 02] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo & Marco Ratto. *Sensitivity Analysis in Practice*. John Wiley & Sons, Ltd, Chichester, UK, feb 2002. [18](#), [37](#)
- [Sanchez 12] Susan M. Sanchez & Hong Wan. *Work smarter, not harder : a tutorial on designing and conducting simulation experiments*. In C. Laroque, J. Himmelspach, R. Pasupath, O. Rose & A. M. Uhrmacher, editeurs, Winter Simulation Conference, pages 1929–1943, Berlin, Germany, 2012. IEEE. [30](#)
- [Saporta 11] Gilbert Saporta. *Probabilités, analyse des données et statistique*. TECHNIP, 2011. [31](#)
- [Sarrafzadeh 06] Maryam Sarrafzadeh, Bill Martin & Afsaneh Hazeri. *LIS professionals and knowledge management : some recent perspectives*. Library Management, vol. 27, no. 9, pages 621–635, dec 2006. [53](#)
- [Schimmerling 17] Paul Schimmerling, Lucie Burçon, M. Sargent, Gaëtan Blondet, Nassim Boudaoud, Julien Le Duigou, Gilles Besombes, Ahmed Hammoumi, Pascal Bonnaud, Bruno Colin, Miguel Dinis, Didier Jeannel, Jean Leung & Laurent Vivet. *FD-X06-081 : Mise en oeuvre des plans d'expériences par essai ou par simulation numérique : choix et exemples d'application*, 2017. [30](#), [41](#)
- [Schoof 94] Larry A Schoof & Victor R Yarberry. *EXODUS II : a finite element data model*. SAND92-2137, Sandia National Laboratories, Albuquerque, NM, vol. 1, 1994. [62](#)

- [Sergent 09] M. Sergent, Delphine Dupuy, Bernard Corre & Magalie Claeys-bruno. *Comparison de méthode criblage pour la simulation numérique*. In 41èmes Journées de Statistique, SFdS, pages 1–6, Bordeaux, France, 2009. [37](#)
- [Shahin 14] Mohamed a. Shahin. *State-of-the-art review of some artificial intelligence applications in pile foundations*. Geoscience Frontiers, pages 1–12, nov 2014. [35](#)
- [Shraim 89] Mustafa S. Shraim. *An expert system for designing statistical experiments*. PhD thesis, Ohio University, 1989. [78](#)
- [Shu-Hsien Liao 05] Shu-Hsien Liao. *Expert system methodologies and applications—a decade review from 1995 to 2004*. Expert Systems with Applications, vol. 28, no. 1, pages 93–103, jan 2005. [56](#), [76](#)
- [Siddiqa 16] Aisha Siddiqa, Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Mohsen Marjani, Shahabuddin Shamshirband, Abdullah Gani & Fariza Nasaruddin. *A survey of big data management : Taxonomy and state-of-the-art*. Journal of Network and Computer Applications, vol. 71, pages 151–166, aug 2016. [14](#)
- [Sirin 07] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur & Yarden Katz. *Pellet : A practical OWL-DL reasoner*. Web Semantics, vol. 5, no. 2, pages 51–53, 2007. [100](#)
- [Slanzi 14] Debora Slanzi & Irene Poli. *Evolutionary Bayesian Network design for high dimensional experiments*. Chemometrics and Intelligent Laboratory Systems, vol. 135, pages 172–182, jul 2014. [82](#)
- [Sluiter 09] R Sluiter. *Interpolation methods for climate data : literature review*. Rapport technique, KNMI, 2009. [36](#)
- [Sobol’ 67] I.M Sobol’. *On the distribution of points in a cube and the approximate evaluation of integrals*. USSR Computational Mathematics and Mathematical Physics, vol. 7, no. 4, pages 86–112, jan 1967. [27](#)
- [Soldatova 06] Larisa N Soldatova & Ross D King. *An ontology of scientific experiments*. Journal of The Royal Society Interface, vol. 3, no. 11, pages 795–803, dec 2006. [71](#), [102](#)
- [Spirtes 93] Peter Spirtes, Clark Glymour & Richard Scheines. Causation, prediction, and search. Springer-Verlag, 1993. [114](#)
- [Stark 04] John Stark. Product lifecycle management, 21st century paradigm for product realisation. Springer, 1st edition, 2004. [12](#)
- [Subba Rao 07] Amara Venkata Subba Rao & Dilip Kumar Pratihar. *Fuzzy logic-based expert system to predict the results of finite element analysis*. Knowledge-Based Systems, vol. 20, no. 1, pages 37–50, 2007. [84](#)
- [Taguchi 00] Genichi Taguchi, Subir Chowdhury & Shin Taguchi. Robust engineering. McGraw-Hill Professional, 2000. [18](#), [26](#)
- [Tasdemir 11] Sakir Tasdemir, Ismail Saritas, Murat Ciniviz & Novruz Allahverdi. *Artificial neural network and fuzzy expert system comparison for prediction of performance and emission parameters on a gasoline engine*. Expert Systems with Applications, vol. 38, no. 11, pages 13912–13923, 2011. [79](#)



- [Terzi 10] Sergio Terzi, Abdelaziz Bouras, Debashi Dutta, Marco Garetti & Dimitris Kiritsis. *Product lifecycle management - From its history to its new role*. International Journal of Product Lifecycle Management, vol. 4, no. 4, page 360, 2010. [12](#), [13](#), [14](#)
- [Urrea 15] Claudio Urrea, Gloria Henríquez & Marcela Jamett. *Development of an expert system to select materials for the main structure of a transfer crane designed for disabled people*. Expert Systems with Applications, vol. 42, no. 1, pages 691–697, 2015. [78](#)
- [Vandewoestyne 06] Bart Vandewoestyne & Ronald Cools. *Good permutations for deterministic scrambled Halton sequences in terms of  $L_2$ -discrepancy*. Journal of Computational and Applied Mathematics, vol. 189, no. 1-2, pages 341–361, may 2006. [27](#)
- [Vanschoren 10] Joaquin Vanschoren & Larisa N Soldatova. *Exposé : An ontology for data mining experiments*. In Proc. of the 3rd Int. Workshop on Third Generation Data Mining : Towards Service-oriented Knowledge Discovery (SoKD), pages 31–46, Barcelona, Spain, 2010. [71](#), [102](#)
- [Vila 15] C. Vila, J.V. Abellán-Nebot, J.C. Albiñana & G. Hernández. *An Approach to Sustainable Product Lifecycle Management (Green PLM)*. The Manufacturing Engineering Society International Conference, MESIC 2015, vol. 132, pages 585–592, 2015. [15](#)
- [Wang 07] G. Gary Wang & S. Shan. *Review of Metamodeling Techniques in Support of Engineering Design Optimization*. Journal of Mechanical Design, vol. 129, no. 4, page 370, 2007. [30](#)
- [Weiner 92] David J. Weiner. *Expert systems to aid in the formulation of hypotheses and design of experiments in biomedical research*. Mathematical and Computer Modelling, vol. 16, no. 6-7, pages 185–198, jun 1992. [78](#)
- [Wiener 38] Norbert Wiener. *The homogeneous chaos*. American Journal of Mathematics, vol. 60, no. 4, pages 897–936, 1938. [34](#)
- [Xiu 02] Dongbin Xiu & George Em Karniadakis. *The Wiener–Askey polynomial chaos for stochastic differential equations*. SIAM journal on scientific computing, vol. 24, no. 2, pages 619–644, 2002. [34](#)



# Annexe A

## Figures annexes

FIGURE A.1 – Logigramme du processus de PEN implémenté

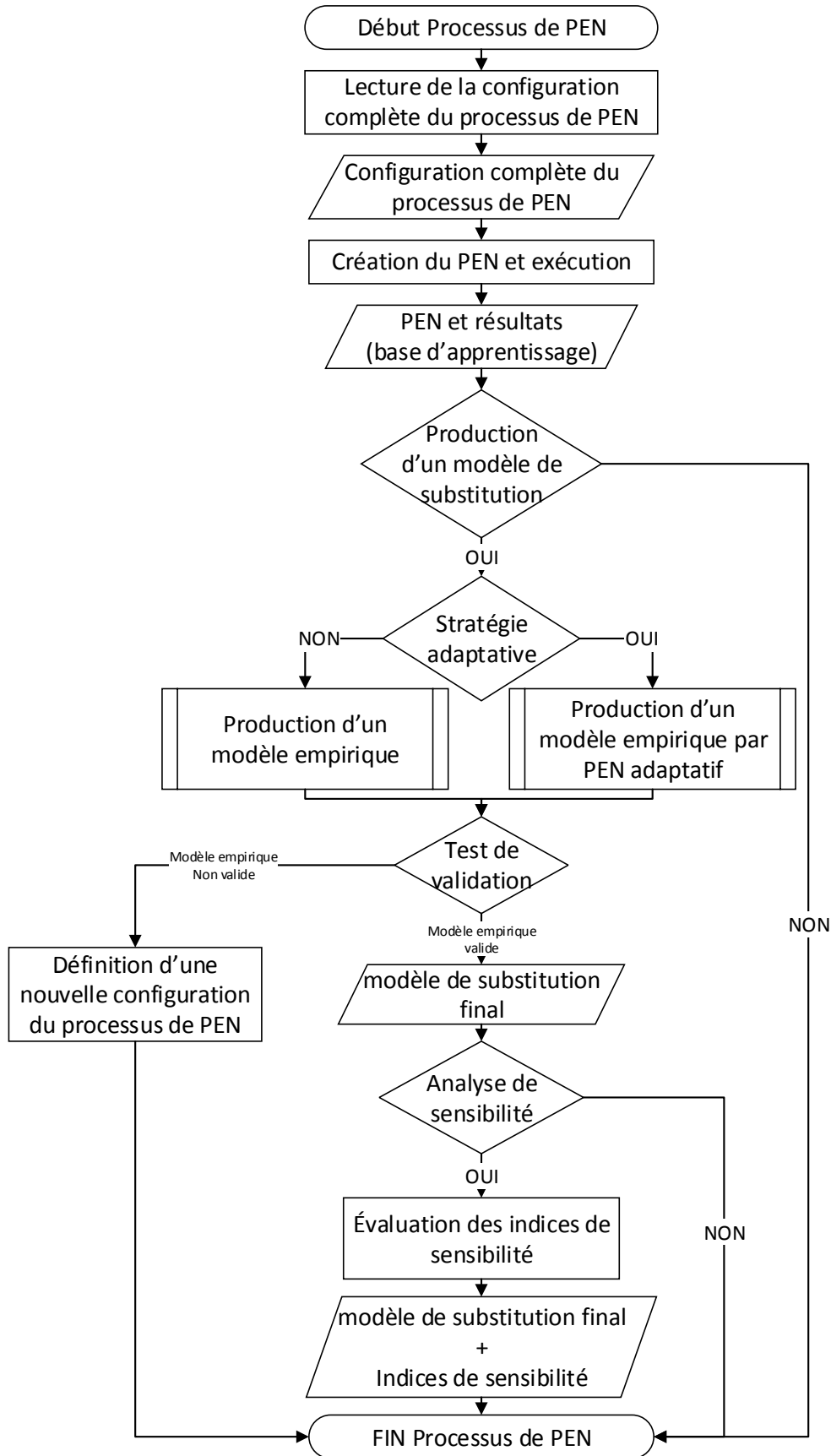
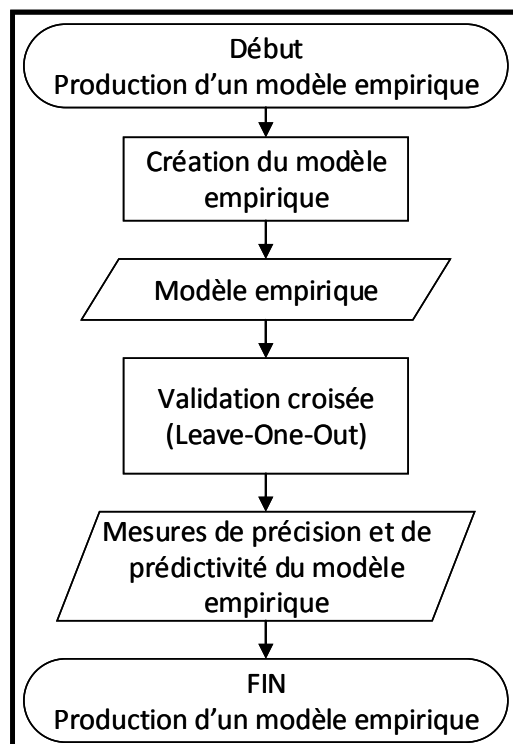
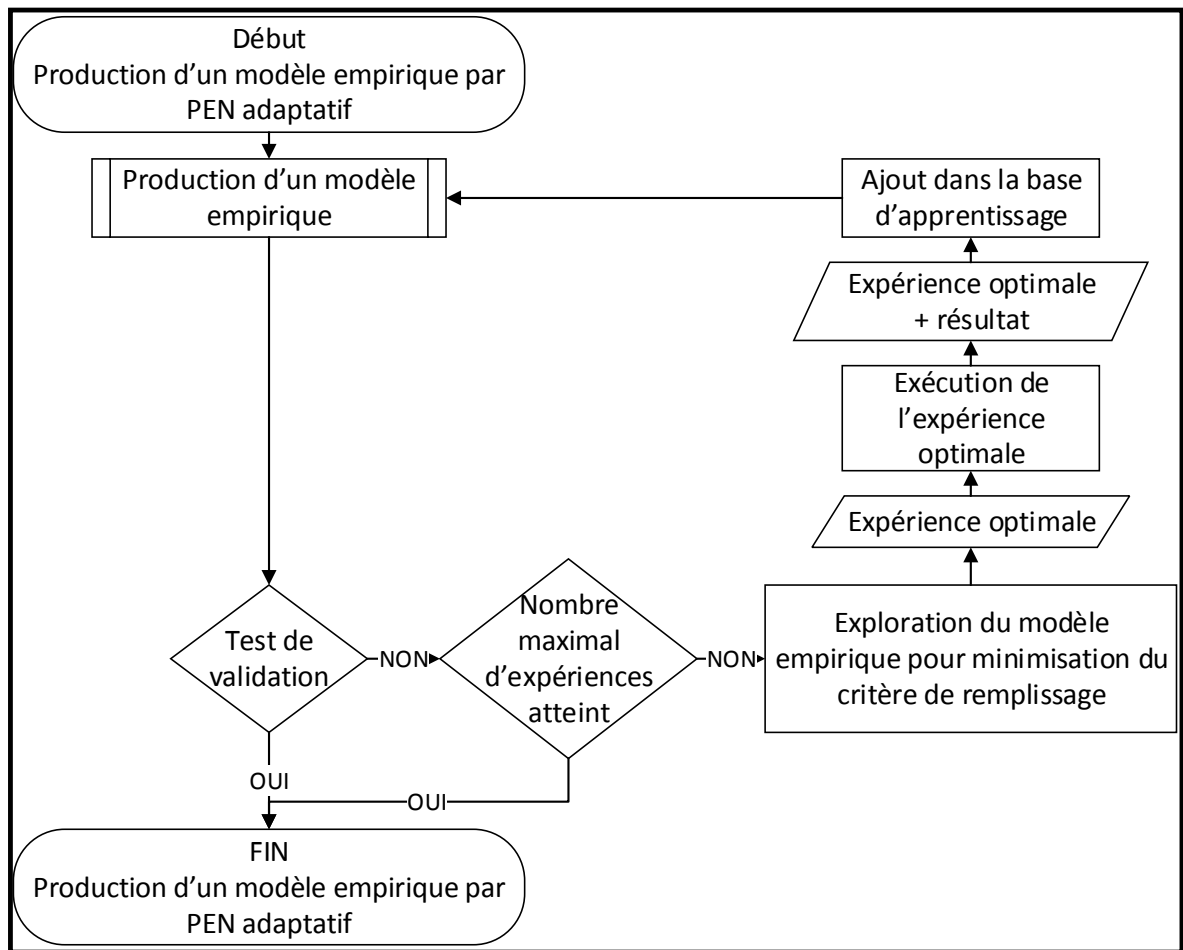


FIGURE A.2 – Logigramme du processus de PEN implémenté pour la production de modèle de substitution



## Résumé

Ces travaux de thèse se placent dans le contexte du développement de produit au sein de l'industrie mécanique. Le besoin de compétitivité des entreprises, dans un contexte économique mondialisé, repose sur l'amélioration de la qualité des produits, la réduction des coûts et la réduction du temps de mise sur le marché. La simulation numérique occupe une place importante dans le processus de conception de produits complexes pour atteindre ces objectifs, mobilisant des expertises diverses dans une entreprise étendue. Les Plans d'Expériences Numériques (PEN) sont de plus en plus utilisés pour intégrer le caractère variable des propriétés du produit et de son environnement directement lors des simulations. Un processus de PEN apporte des méthodes de planification et d'analyse d'un ensemble de simulations, pour mieux maîtriser les performances du produit. La problématique traitée repose sur deux points principaux. D'une part, la définition d'un processus de PEN est caractérisée par de nombreux choix et l'utilisation de méthodes complexes, nécessitant une expertise avancée. Cette définition est d'autant plus complexe que le modèle de simulation est complexe et coûteux à exécuter. D'autre part, l'utilisation de PEN conduit à une production de grands volumes de données hétérogènes, par l'exécution intensive de simulations numériques. Ces travaux portent sur l'obtention rapide de la configuration optimale du processus de PEN. Ainsi, le gain de temps se place à la fois durant la préparation de la démarche et durant son exécution. Ces travaux se sont orientés vers la réutilisation des connaissances en entreprise pour un système à base de connaissances, composé d'une ontologie spécifique, pour supporter la capitalisation et le partage des connaissances liées aux PEN, et d'un moteur d'inférences, basé sur les réseaux bayésiens, réutilisant les données et les connaissances pour la prédiction de configurations efficaces et innovantes de processus de PEN. Cette proposition est illustrée par une application sur un produit industriel issue du secteur automobile.

**Mots-clés : Plan d'expériences numériques, Ontologie, Réseau bayésien, Système à base de connaissances**

## Abstract

In order to improve industrial competitiveness, product design relies more and more on numerical tools, such as numerical simulation, to develop better and cheaper products faster. Numerical Design of Experiments (NDOE) are more and more used to include variabilities during simulation processes, to design more robust, reliable and optimized product earlier in the product development process. Nevertheless, a NDOE process may be too expensive to be applied to a complex product, because of the high computational cost of the model and the high number of required experiments. Several methods exist to decrease this computational cost, but they required expert knowledge to be efficiently applied. In addition to that, NDoE process produces a large amount of data which must be managed. The aim of this research is to propose a solution to define, as fast as possible, an efficient NDoE process, which produce as much useful information as possible with a minimal number of simulations, for complex products. The objective is to shorten both process definition and execution steps. A knowledge-based system is proposed, based on a specific ontology and a bayesian network, to capitalise, share and reuse knowledge and data to predict the best NDoE process definition regarding to a new product. This system is validated on a product from automotive industry.

**Keywords: Numerical design of experiments, Ontology, Bayesian Network, Knowledge-based system**