GRANDE CULTURE GCHP2E à Hautes Performances Economiques et Environnementales

Etude de faisabilité du développement et de la valorisation d'une base de données sur l'évolution des pressions biotiques dans les parcelles agricoles

Document de présentation, novembre 2011

INRA

INSTITUT TECHNIQUE DE LA BETTERAVE ITB

ARVALIS Institut du végétal

DRAAF Midi-Pyrénées

CETIOM Centre Technique Interprofessionnel des Oléagineux Métropolitains

Institut Technique du Lin

PROLEA UNIP

AGRICULTURES & TERRITOIRES CHAMBRES D'AGRICULTURE

« Etude de faisabilité du développement et de la valorisation d'une base de données sur l'évolution des pressions biotiques dans les parcelles agricoles»

I. Contexte

Le GIS Grande Culture à Hautes Performances Economiques et Environnementales (GC-HP2E) finance actuellement une étude de faisabilité du **développement** et de la **valorisation d'une base de donnée sur l'évolution des pressions biotiques dans les parcelles agricoles**. Le projet, coordonné par le réseau Protection Intégrée des Cultures (PIC) de l'INRA et regroupant des partenaires de la recherche et du développement agricole (INRA, ITB, Arvalis-Institut du Végétal, CETIOM, ITL, UNIP, APCA, SRAl Midi-Pyrénées), s'inscrit dans les objectifs du Grenelle de l'environnement, et en particulier celui qui vise à réduire de 50% l'usage des produits phytosanitaires d'ici 2018 si possible. Il part du constat que **les niveaux de pressions biotiques subies par les différentes productions végétales sont difficiles à connaître au niveau national** (Aubertot et al., 2005[1]).

Or, pendant plusieurs décennies, les **Services Régionaux de la Protection des Végétaux** (SRPV) ont produit des documents destinés à tenir informés les agriculteurs et leurs conseillers des pressions biotiques subies par les différentes cultures : **avertissements agricoles, bilans de campagne régionaux et nationaux** .
L'ensemble des ces données historiques de la Protection des Végétaux (PV) représentent une source d'informations extrêmement précieuse, d'autant plus qu'elles recouvrent **une très large gamme de situations de production, sur des séries temporelles longues et sur l'ensemble du territoire national**.

II. Valorisations envisagées des données

Cette démarche permettra de **rassembler et** de **valoriser le patrimoine incomparable constitué au sein des SRPV en plus de six décennies** et qui risque de se perdre avec l'évolution des Services Régionaux de l'Alimentation (SRAl). Les données d'autres organismes (CA, ICTA) pourront également être sauvegardées et partagées grâce à l'outil.

Les **domaines d'utilisation scientifiques et techniques de ces données**, identifiés à partir de consultations d'acteurs de la recherche-développement et d'exemples existants dans la littérature, sont **très nombreux**. On peut citer notamment les valorisations potentielles suivantes :

- Une meilleure description des évolutions spatio-temporelles des populations de bioagresseurs et/ou des dégâts à différentes échelles. Ceci permettrait notamment une meilleure hiérarchisation des pertes de rendement effectivement engendrées par les bioagresseurs.
- L'analyse des relations entre pratiques agricoles et pressions biotiques afin de faciliter la mise au point de stratégie de protection intégrée des cultures.
- L'analyse des relations dégâts-dommages à partir de résultats expérimentaux en parcelle comportant des zones traitées et non traitées.
- Le développement de modèles et/ou d'outils d'aide à la décision pour le raisonnement des applications de produits phytosanitaires, et pour la conception de stratégies de protection/ production intégrée
- L'analyse des relations entre variables climatiques et pressions biotiques afin d'imaginer des stratégies d'adaptation des systèmes de culture au changement climatique.
- La réflexion méthodologique sur l'optimisation des systèmes actuels de suivi de bioagresseurs.

[1]Aubertot J.N., J.M. Barbier, A. Carpentier, J.J. Gril, L. Guichard, P. Lucas, S. Savary, I. Savini, M. Voltz (éditeurs), 2005. *Pesticides, agriculture et environnement. Réduire l'utilisation des pesticides et limiter leurs impacts environnementaux*. Rapport d'Expertise scientifique collective, INRA et CEMAGREF (France).

III. Etapes d'un projet de création de base de données

1. Collecte et numérisation des documents

a. Choix des documents

Pour des raisons à la fois scientifiques (intérêt du contenu) et pratiques (facilité de récupération, non redondance entre les données qu'ils contiennent) **nous nous intéresserons en priorité à deux types de documents de la PV : les avertissements agricoles et les bilans de campagne nationaux par thématique**. Pourront s'y ajouter par la suite les bilans de campagne régionaux et éventuellement les fiches de notation de la PV. **Nous chercherons aussi à intégrer dans l'outil des données d'autres organismes comme les ICTA et les CA, dont une partie a été inventoriée**. Les Bulletins de Santé du Végétal (en libre accès sur Internet) seront également rassemblés, afin d'assurer la continuité entre les données anciennes et les données récentes.

La volumétrie totale des deux types de documents de la PV cités ci-dessus est estimée à *ca.* 200 000 pages.

La démarche présentée ci-dessous permet de rassembler l'ensemble des documents de la PV concernant les grandes cultures. En option, nous proposons une numérisation de fonds des Archives Nationales, plus anciens, qui permettrait d'élargir notre collection aux autres filières (le système d'avertissements agricoles a commencé avec l'arboriculture et la vigne, puis s'est étendu peu à peu aux grandes cultures).

b. Droit d'exploitation des documents

Les droits sur les documents de la PV sont détenus par la Direction Générale de l'Alimentation (DGAL), dont l'autorisation est donc indispensable pour récupérer et diffuser ces collections. Une première réunion a eu lieu en juin 2011 entre l'équipe de projet et la DGAL. A la demande de cette dernière, **un argumentaire mettant en avant l'intérêt du projet et destiné aux services régionaux a été rédigé par l'équipe de projet. Il sera diffusé aux SRAl**, afin qu'ils soient informés de la démarche et puissent faciliter l'accès de leurs archives au groupe de projet. Un **questionnaire sur les types de documents disponibles dans les SRAl** a également été crée par l'équipe de projet et transmis à la DGAL pour qu'elle le communique aux services régionaux.

Pour permettre le montage du projet suite à l'étude de faisabilité, les échanges décrits ci-dessus devront être formalisés : une autorisation écrite de la DGAL à accéder aux documents et à les exploiter est indispensable au démarrage du projet.

c. Collecte et numérisation des documents

Les délais de numérisation entrevus pour la collection d'avertissements agricoles présents à la Bibliothèque Nationale de France (BnF) nous incitent à proposer **une collecte et une numérisation des documents en 2 temps :**

- Collecte de documents existants dans les SRAl, les ICTA et les CA

i. Documents sous format électronique

Pour les données des SRAl, la première phase du projet consistera à collecter les documents existant sous format électronique (généralement depuis 1995) afin d'éviter une phase de numérisation inutile. Une transformation du format de ces documents sera néanmoins

nécessaire afin qu'ils puissent être exploités de la même façon que ceux issus de la numérisation. Cette phase se fera grâce à des échanges avec les agents des différents SRAl, avec l'accord de la DGAL.
La même démarche sera effectuée pour les documents des ICTA et des CA existant sous format électronique.
Nous collecterons dès le début du projet l'ensemble des BSV produits par les différentes régions depuis 2009, et disponibles en ligne sous format pdf sur les sites des Directions Régionales de l'Alimentation, de l'Agriculture et de la Forêt (DRAAF).

ii. Documents sous format papier

Les documents antérieurs à 1995 et n'existant que sous format papier devront être numérisés et soumis à une reconnaissance optique de caractères (OCR) par un prestataire. Nous proposons pour les bilans de campagne annuels, non présents à la BnF, une collecte des documents dans les SRAl. Comme chacun d'entre eux recevait l'ensemble des bilans nationaux, nous identifierons grâce à des enquêtes 2 ou 3 services disposant d'archives importantes, et ferons numériser les documents de ces régions.
Des documents des ICTA et des CA seront numérisés/OCRisés suite à un inventaire détaillé des fonds, une évaluation de leur adéquation avec les objectifs du projet et une analyse des modalités de leur mise à disposition.

La numérisation/OCRisation de tous les documents sera réalisée par des sociétés spécialisées selon un cahier des charges rédigé par nos soins et pourra faire l'objet d'un appel d'offres selon la réglementation des marchés publics. Les différentes options techniques (choix du format de sortie, de la résolution, du type d'OCR…) ont été étudiées avec l'aide de spécialistes.

- Numérisation des avertissements agricoles de la BnF

En ce qui concerne les avertissements agricoles, il en existe une collection conséquente (50 années d'avertissements pour presque toutes les régions de France hors outre-mer, à partir du début de la décennie 1960) **à la Bibliothèque nationale de France** (BnF), en vertu du dépôt légal. Compte tenu de l'ampleur de ce fond (très difficile à égaler en rassemblant les archives d'avertissements de chaque région) et de la centralisation de ces différents documents régionaux en un même lieu, la numérisation des collections d'avertissements agricoles de la BnF nous semble la solution la plus pertinente pour ce type de document. Cette opération s'inscrira dans les marchés de numérisation de la BnF et devra fera l'objet d'un accord préalable de la direction de celle-ci, suite à un dépôt de dossier par l'équipe de projet. **En raison du planning de numérisation très chargé de la BnF, cette action ne pourra pas se faire avant 2013.** Par contre, une action d'inventaire précis de tous les documents présents à la BnF et donc des lacunes de ce fonds, sera entreprise dès le début du projet afin de pouvoir combler les manques à l'aide des documents éventuellement présents dans les SRAl.

La numérisation à la BnF s'effectuera selon les cahiers des charges établies par celle-ci, et pourra si besoin être suivie d'un retraitement des documents si le taux de reconnaissance des caractères ne semble pas adapté aux besoins du projet. La négociation avec la BnF devra inclure la mise à disposition des fichiers sources (Alto OCR et fichiers image).

La numérisation à la BnF s'effectuera selon les cahiers des charges établies par celle-ci, et pourra si besoin être suivie d'un retraitement des documents si le taux de reconnaissance des caractères ne semble pas adapté aux besoins du projet. La négociation avec la BnF devra inclure la mise à disposition des fichiers sources (Alto OCR et fichiers image).

Bien que cette phase de numérisation soit la dernière à être mise en œuvre, elle demande une préparation dès le début du projet afin de s'assurer de l'accord de la BnF et appréhender les modalités pratiques du procédé (coûts, conditions de fourniture des fichiers source…).

- **Option :** *numérisation des avertissements agricoles des Archives nationales*

Des fonds anciens d'avertissements agricoles (années 1940-1980) et de rapports nationaux (années 1940-milieu de la décennie 1970) sont disponibles aux Archives nationales à Fontainebleau. Les documents les plus anciens, non redondants avec ceux de la BnF, concernent essentiellement d'autres filières que les grandes cultures (arboriculture, vigne …, à savoir les seules filières suivies au début du système d'avertissements). Leur numérisation et OCRisation pourrait être réalisée sur place ou chez un prestataire sans contraintes particulières et selon notre cahier des charges. Il pourrait donc être intéressant, si l'on souhaite élargir l'étude à d'autres cultures, d'inclure ce fond dans la démarche de numérisation. De plus, pour les séries les plus récentes (existant à la fois à la BnF et aux Archives nationales), il serait possible de combler certaines lacunes des fonds de la BnF grâce aux collections des Archives nationales
En raison du déménagement des Archives Nationales au cours de la 2ième moitié de 2012, la numérisation de ces fonds devrait avoir lieu avant juillet 2012 ou à partir de 2013.

2. Mise à disposition des documents

a. Schéma de mise à disposition envisagé

La base de données comportera d'une part des **documents indexés par mots-clés** pour faciliter l'exploitation des productions issues du savoir-faire des différents SRPV, et d'autre part de **séries de données chiffrées** permettant la caractérisation de l'évolution des pressions biotiques sur le long terme.
Nous souhaitons mettre à disposition le plus rapidement possible le corpus de données textuelles, à savoir dans un premier temps les BSV, les avertissements agricoles et les bilans de campagne nationaux de la PV ainsi que les rapports des ICTA et des CA sous format pdf (fichiers électroniques récents récupérés directement et documents anciens numérisés). Ces documents seront indexés par mots-clés et découpés en sections thématiques. L'utilisateur pourra effectuer une **recherche multi-critères** (par culture, bioagresseur, région, année…) sur ce corpus, et accéder ainsi aux documents ou aux sections de documents abordant la thématique de sa requête. S'il **s'intéresse à des séries de données chiffrées particulières et souhaite les extraire** pour les exploiter, il pourra les saisir lui-même dans la base de données, où elles seront alors accessibles à l'ensemble des utilisateurs. Cette **démarche participative** permettra de rendre immédiatement disponibles les documents dans leur intégralité, et de **cibler les données chiffrées plus détaillées à extraire en fonction des besoins réels des utilisateurs**.

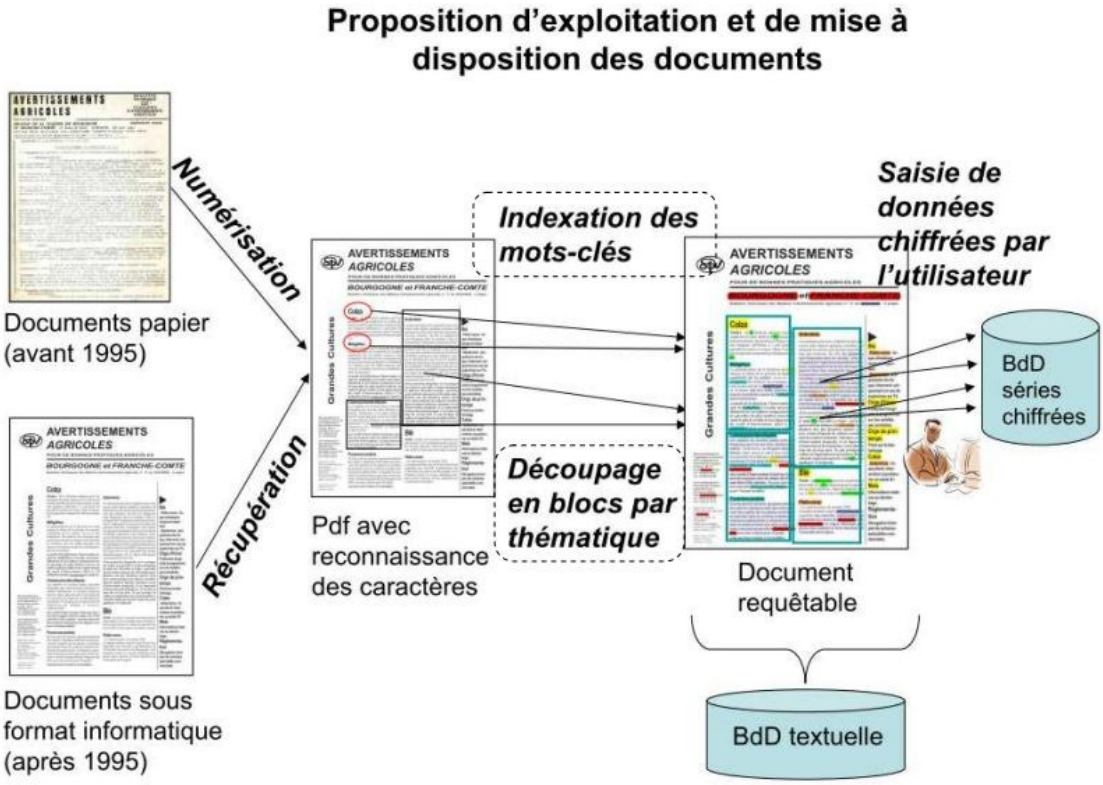

Figure 1. Proposition de mise à disposition des documents.

b. Les outils envisagés

Les données textuelles seront mises à disposition dans un outil de recherche sémantique. Nous travaillons actuellement avec 3 équipes de l'INRA spécialisées dans ce domaine (équipe CoreTexT dirigée par Marc Barbier, équipe d'Odile Hologne de la DV/IST et l'équipe Bibliome dirigée par Claire Nédellec) sur le développement d'un tel outil, à partir d'un échantillon de documents représentatifs.

En ce qui concerne les **séries de données chiffrées**, nous sommes en train d'étudier l'opportunité de les intégrer dans la **base de données centrale en cours de construction par la DGAL (Epiphyt)**, afin d'assurer la continuité entre données passées et actuelles. Nous avons échangé à ce sujet avec la responsable informatique d'Epiphyt.

Des fonctionnalités indispensables aux outils ont été proposées par de futurs utilisateurs potentiels lors d'entretiens. L'accès aux outils se fera à travers un portail spécifique, qui pourra aussi comporter des informations utiles pour l'exploitation des données (suivi des projets de valorisation en cours, liens vers des bases de données connexes intéressantes…).

3. Accompagnement de projets pilotes

Lors d'entretiens réalisés avec de futurs utilisateurs potentiels, nous avons identifié un certain nombre de projets de valorisation des données, textuelles ou chiffrées. Nous sélectionnerons parmi ceux-ci **quelques projets pilotes diversifiés et pouvant démarrer rapidement**. Les porteurs de ces projets seront fortement impliqués dès le début dans la conception de la base de données, au sein d'un comité utilisateurs qui veillera à ce que la base corresponde aux besoins exprimés. En retour, ces porteurs de projet bénéficieront de la part de l'équipe de projet d'un accompagnement dans l'exploration des documents et la constitution de séries de données. Ces projets pilotes, qui montreront la diversité des questions techniques et scientifiques à laquelle la base de données peut répondre, serviront aussi de « vitrine » à l'outil.

Le tableau ci-dessous recense quelques projets de valorisation variés recensés auprès de futurs utilisateurs potentiels :
De plus, les porteurs de deux projets (Rédupest, projet INRA soutenu par le GIS GC-HP2E et un projet du GEVES) souhaitent disposer très rapidement de données pour des études en cours.

Thématique	**Organisme porteur**
Construction d'un OAD sur l'anthracnose du pois	UNIP
Analyse et hiérarchisation des dommages causés par les bioagresseurs du blé à l'échelle nationale	INRA (UMR AGIR)
Analyse des relations entre occupation des terres et intensité de bioagresseurs à l'échelle de petites régions	INRA (UMR Agronomie)
Etude de la co-évolution de bioagresseurs par analyse textuelle	INRA (UMR BGA)
Etude des savoirs et dispositifs des transitions vers une production agricole durable (passage avertissements - BSV)	INRA (UMR SenS)
- Reconstitution du climat passé en utilisant des vieilles séries d'observation phénologiques - Modélisation des effets du changement climatique sur l'évolution des bioagresseurs des principales cultures à l'échelle nationale	INRA (US Agroclim)

4. Réflexion sur le suivi épidémiologique actuel

L'étude des données historiques de la Protection des Végétaux permet d'amorcer un travail de **réflexion sur le suivi épidémiologique actuel** servant à produire les BSV. L'analyse des lacunes constatées dans les données de la PV (données connexes manquantes, absence d'homogénéité dans les notations…) fournit des pistes de travail pour optimiser les observations réalisées actuellement, en ce qui concerne l'échantillonnage des parcelles, le choix des variables recensées et la traçabilité du notateur notamment.
Nous formulerons des **préconisations** à ce sujet **dans le rapport final** de l'étude de faisabilité, qui pourront être suivies par des échanges avec les acteurs des Comités régionaux ou nationaux d'épidémiosurveillance.

III. Montage du projet

1. Planning prévisionnel

Les différentes étapes du projet sont représentés de manière simplifiée ci-dessous :

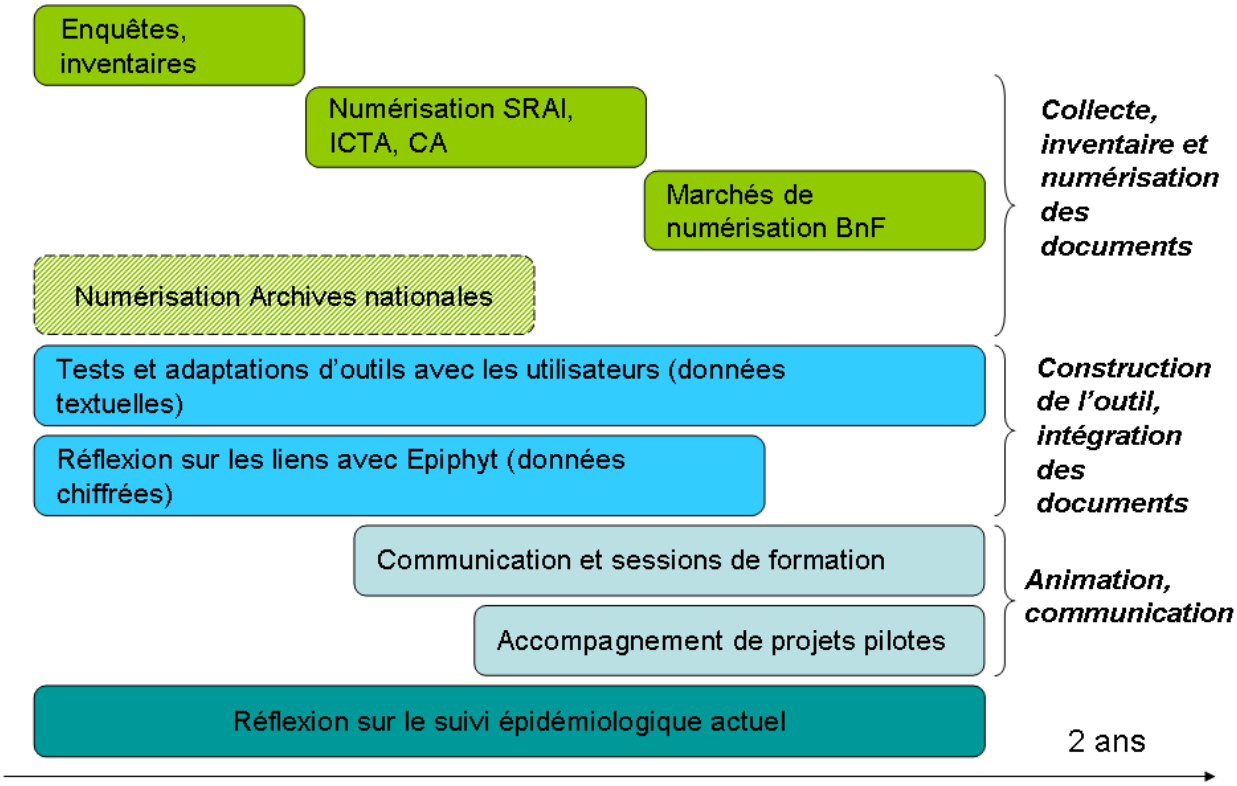

Fig. 2 : déroulement prévisionnel du projet

Nous estimons que **2 ans** devraient suffire à récolter les documents, construire la base de données et mettre en œuvre les premiers projets pilotes. Un suivi du travail à la suite de ces 2 années (gestion de la base de données, accompagnement des projets de valorisation…) sera nécessaire, mais ne demandera pas un travail à plein temps.

2. Moyens humains

La coordination du projet sera assurée par **une personne permanente** d'un organisme faisant partie de l'équipe de projet, qui y consacrera une partie conséquente de son temps de travail, et par **un CDD embauché sur 2 ans à plein temps**. **Un contractuel pourra être embauché sur quelques mois pour seconder le CDD** dans l'inventaire des documents, tâche très coûteuse en temps et plus facile à exécuter à deux.

Le projet sera porté par des représentants des organismes de la recherche et du développement agricole. Un **comité de pilotage** validera les grandes orientations du projet au fur et à mesure de son avancement.

Conclusion :

L'étude dont les résultats sont présentés ici démontre la pertinence d'une base de données sur les pressions biotiques pour répondre à une large diversité de questions scientifiques et techniques exprimées par les acteurs de la recherche et du développement agricoles. Des propositions ont été formulées quant au montage d'un projet de création de cet outil, intégrant à la fois des aspects pratiques de collecte et de numérisation des documents, la construction de l'outil lui-même en interaction avec les utilisateurs et des actions d'animation et d'accompagnement des projets qui en bénéficieront.

Le projet ne pourra être lancé sans une autorisation formelle de la DGAL.

Un démarrage rapide du projet à l'issue de l'étude de faisabilité est souhaitable pour plusieurs raisons :

- Nos investigations dans les services régionaux de quelques régions cas d'étude ont montré la nécessité de **recueillir rapidement les fonds dont ils disposent pour éviter leur destruction lors de déménagements futurs**. De telles situations ont pu survenir dans le passé et expliquent la difficulté à retrouver certains des documents cherchés.

- La communication autour du projet et l'implication de différents acteurs dans l'étude de faisabilité (réponses au questionnaire, participation aux entretiens, construction de prototypes d'outils de mise à disposition des données, conseils sur la numérisation…) ont créé **une dynamique qui risquerait de retomber si le délai entre la fin de l'étude de faisabilité et le lancement du projet est trop important**. Ce travail serait alors à refaire.

- Quelques projets pilotes de valorisation des données ont été identifiés. Là aussi, un démarrage rapide du projet de création de base de données éviterait un essoufflement et une perte d'implication des porteurs de ces projets pilotes. En ce qui concerne les projet Rédupest et le projet du GEVES, l'accès aux données de la PV serait nécessaire avant même la création de la base de données, pour des raisons d'échéances.

N.B : *Cette étude a été réalisée dans le cadre du GIS GC-HP2E et ses résultats ne concernent donc que la filière des grandes cultures, même si certains documents sont communs à d'autres filières.* ***Il semble évident que la démarche pourrait être étendue à ces autres filières, notamment en intégrant dans la démarche les documents des Archives nationales, comme il a été indiqué en option.***

Pour toute question sur ce document, merci de contacter Vincent Cellier (INRA, réseau Protection Intégrée des Cultures) : vincent.cellier@epoisses.inra.fr