Vers la gestion et le partage des données de la recherche : chercheurs et documentalistes scientifiques des évolutions parallèles

Thierry Beguiristain, LIEC,OTELo

M-Christine Jacquemot-Perbal, Inist-CNRS

FréDoc 2015

Un partenariat Inist-OTELo dans un contexte de science ouverte et science des données

- Inist
  - Evolution des activités
    - Gestion et valorisation des données de recherche
  - Adaptation aux besoins des chercheurs
  - Accompagnement des chercheurs

- OTELo

  (Observatoire Terre et Environnement de Lorraine)
  - Fédération de 4 UMR et 1 UPR

    (435 personnes)
  - Interdisciplinarité

    (planètes, géologie, biogéochimie, ecotoxicologie…)
  - Gestion et réutilisation des données
  - Anticipation :

    Science des données/ouverture

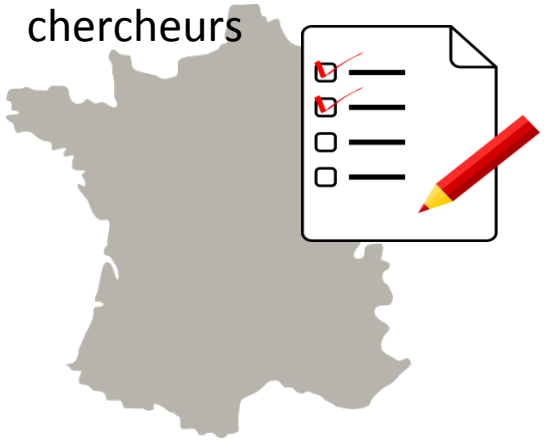
Etude de faisabilité

« Gestion et valorisation des données de recherche »

Paysage des politiques de données

Réunions – entretiens avec les chercheurs

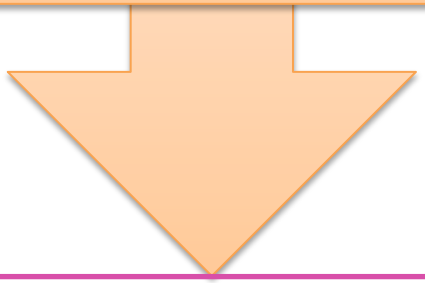
Enquête auprès des chercheurs



- Laboratoires et projets multidisciplinaires
- Gestion d'une grande masse de données (comparaison des résultats, formulation et validation des hypothèses)
- Mutualisation, échange, partage des données
- Utilisation de bases de données (existantes ou à concevoir)
- Outils: espaces collaboratifs, espaces de stockage et de conservation des données

Stratégie adoptée par l'Inist

- Développer une offre de service en réponse aux besoins exprimés par les chercheurs
- S'inscrire dans la mouvance Open Data
  - ➢ Promotion pour l'adoption de bonnes pratiques de gestion, partage et citation des données
  - ➢ Formation/conseil sur les standards, identifiants pérennes, aspects éthiques et juridiques



Redéploiement d'une partie du personnel

- ➢ Documentalistes scientifiques impliqués dans la production de Pascal et Francis

Documentalistes scientifiques : des connaissances et des compétences à construire

Documentalistes scientifiques

Connaissances

- ✓ Culture informationnelle
- ✓ Discipline scientifique
- ✓ Standards de métadonnées bibliographiques
- ✓ Aspects éthiques et juridiques / Publications

Compétences

- ✓ Indexation
- ✓ Terminologie
- ✓ Métadonnées

« Data librarians »

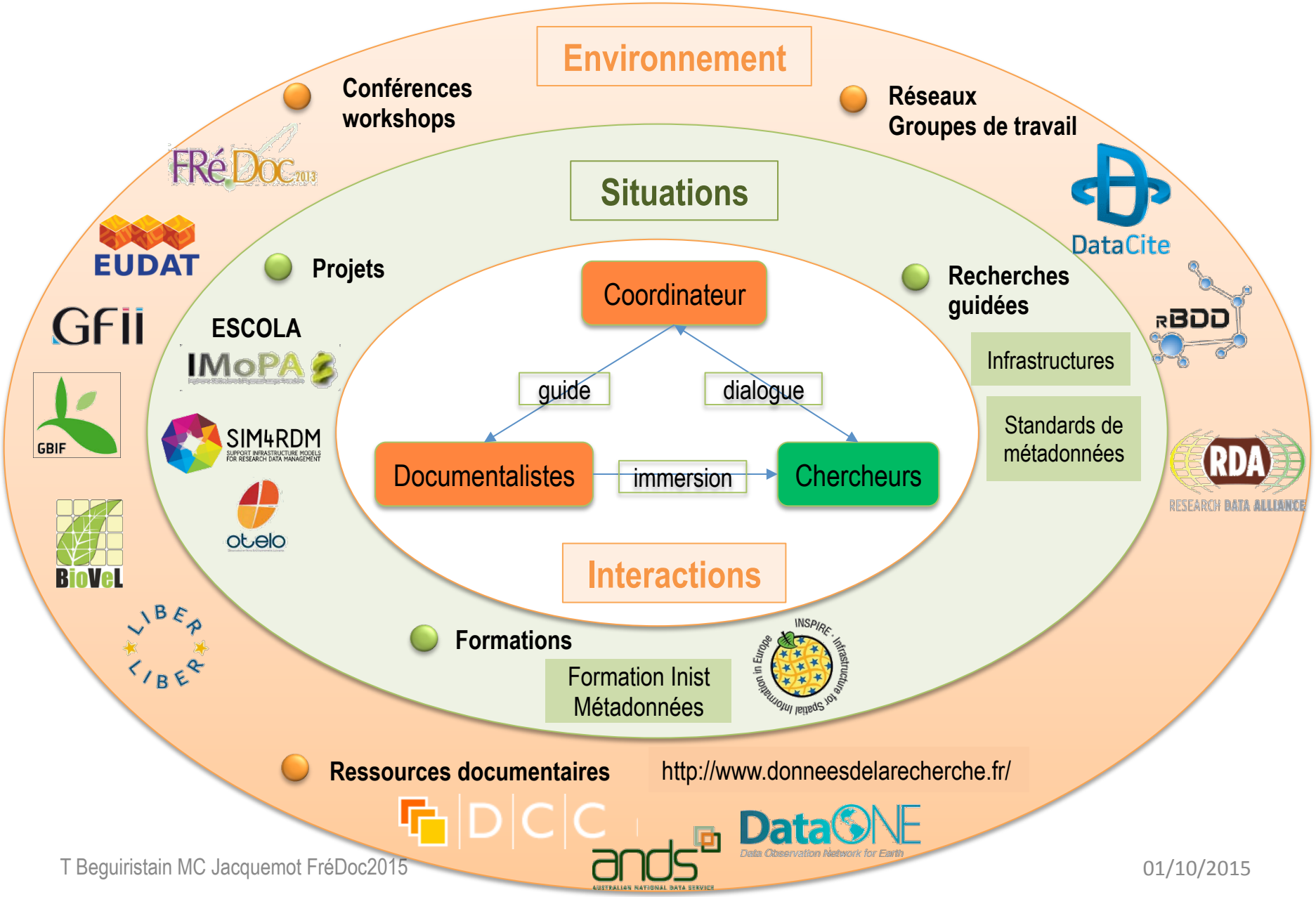
Connaissances

- ✓ Culture des données
- ✓ Compréhension des processus de recherche
- ✓ Standards de métadonnées spécifiques des disciplines
- ✓ Identifiants pérennes
- ✓ Aspects éthiques et juridiques / Données

Compétences

- ✓ Communication
- ✓ Pédagogie
- ✓ Curation des métadonnées
- ✓ Ontologies
- ✓ Informatique : base de données, langages informatiques
- ✓ Web sémantique

Des méthodes actives de formation



Chercheurs: des connaissances et des compétences à construire

Métier des Chercheurs

- ✓ Soucis de la qualité des données (reproductibilité, intégrité)
- ✓ Gestion, analyse et partage de grandes quantités de données
- ✓ Nécessité d'ouverture et de conservation des données
- ✓ Fonds publics = vocation à l'ouverture
- ✓ Verrous juridiques et technologiques qui font obstacle à l'ouverture

« Science des données et de partage des données »

Interrogations

- ✓ Engagements internationaux des institutions?
- ✓ Statut juridique des données vigilance des chercheurs?

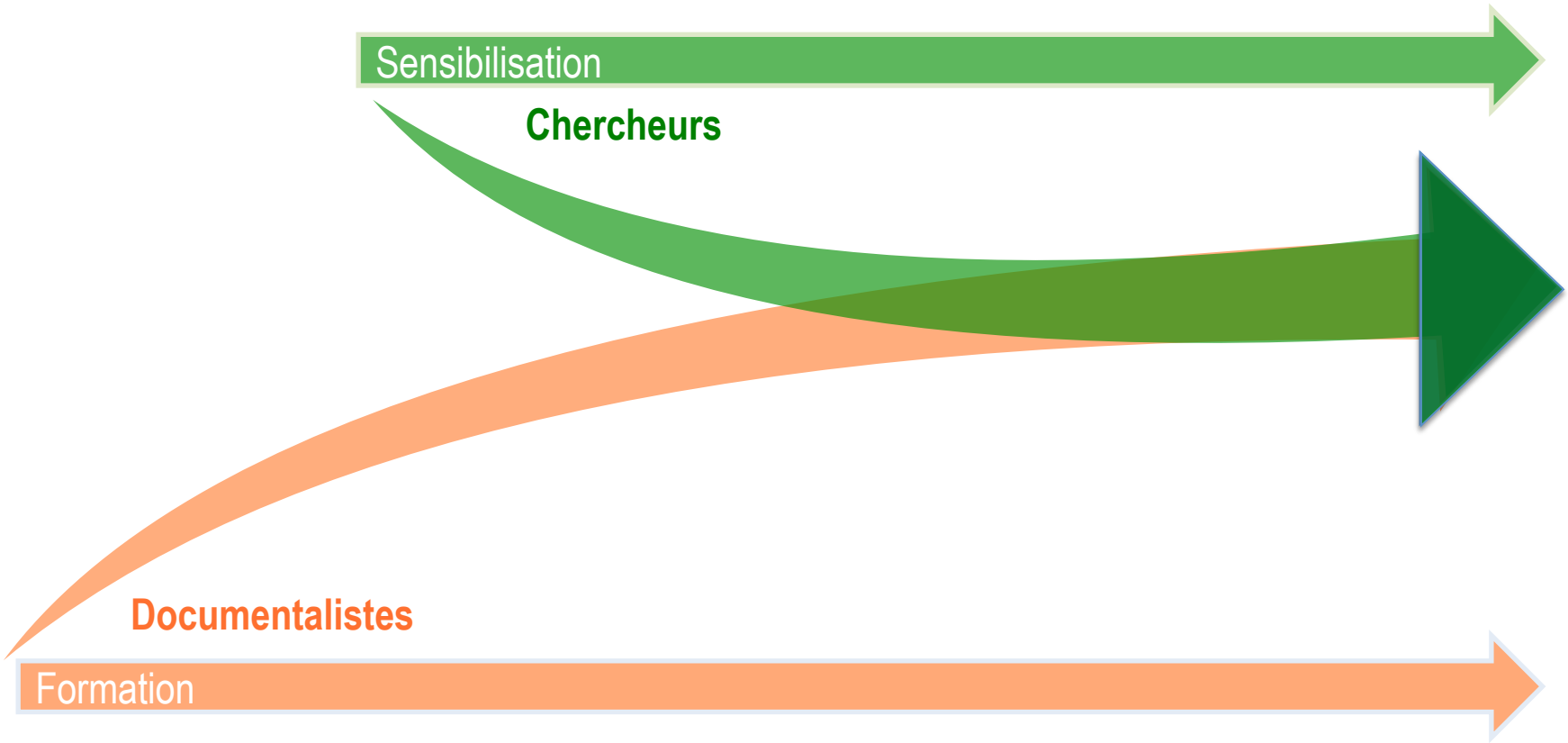
Compétences

- ✓ Gestion et exploitation des données : utilisation des bases de données, outils d'analyse

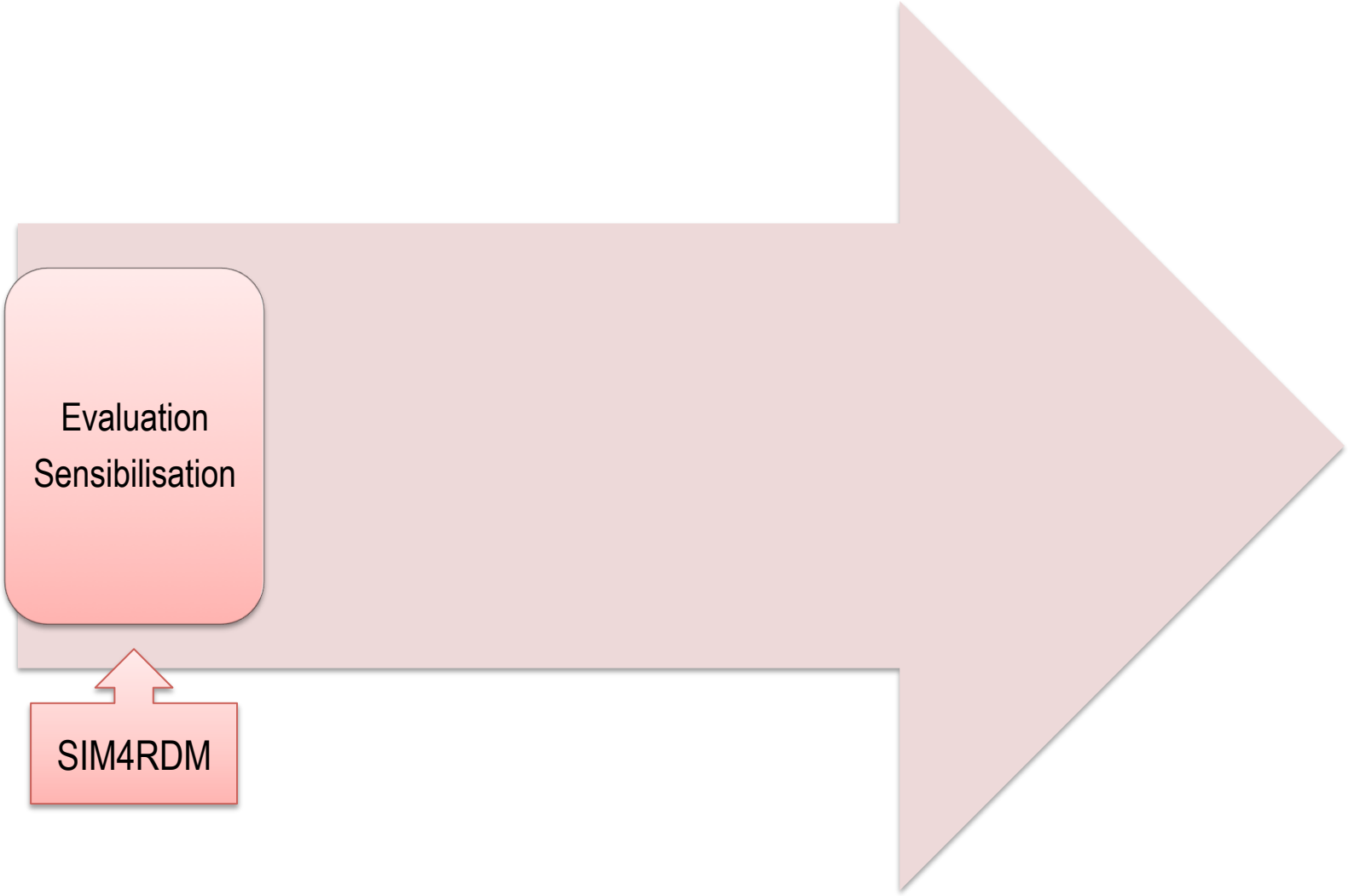
Obstacles

- ✓ Gestion du temps : besoin d'informations digérées, synthétiques
- ✓ Ressources humaines
- ✓ Valorisation du travail

Documentalistes et chercheurs : des évolutions parallèles catalysées par les interactions



Stratégie d'OTELo



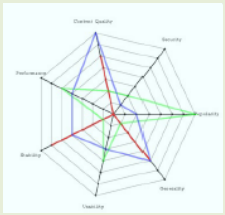
Sensibilisation : évaluation

SIM4RDM
SUPPORT INFRASTRUCTURE MODELS FOR RESEARCH DATA MANAGEMENT
http://www/sim4rdm.eu

« Institution »
OTELo

CHERCHEURS

Questionnaire



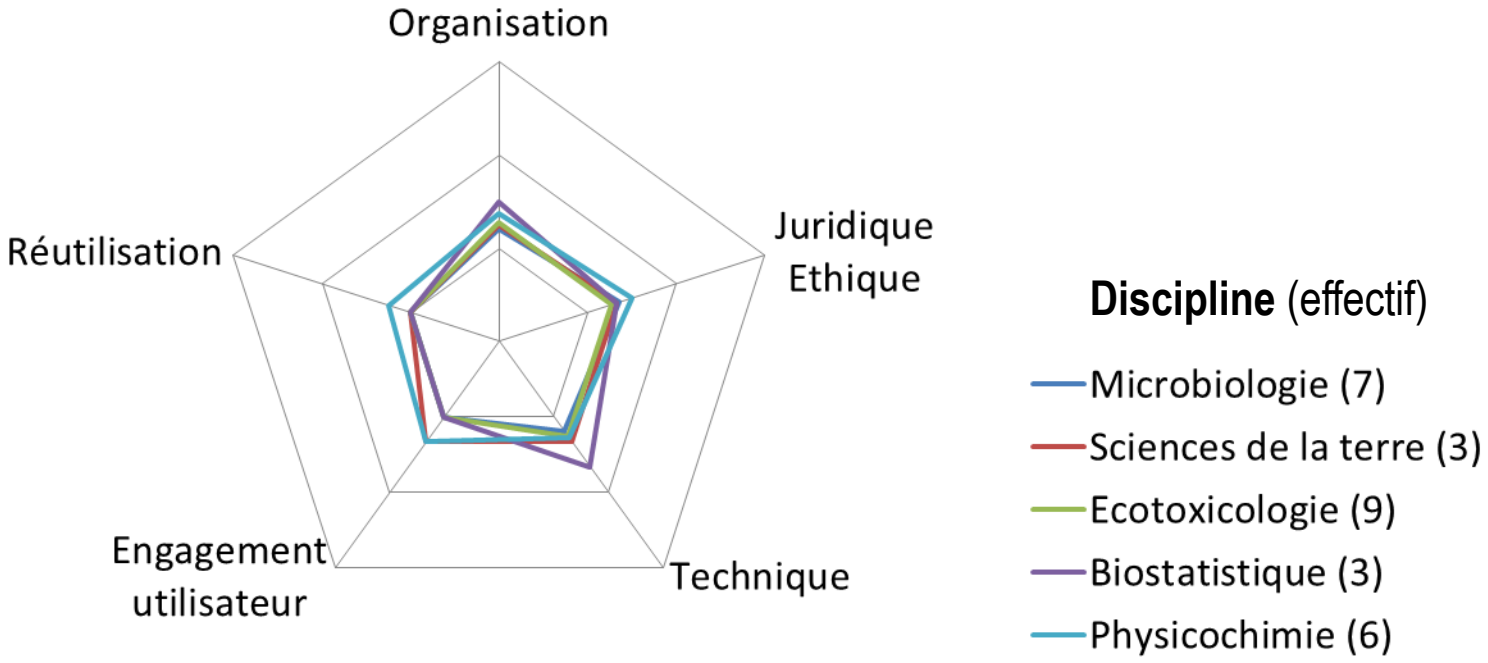
Niveau de maturité

« Infrastructure de soutien »
Inist-CNRS
(documentalistes scientifiques motivés)

Evaluation : bilan

SIM4RDM
SUPPORT INFRASTRUCTURE MODELS FOR RESEARCH DATA MANAGEMENT
http://www/sim4rdm.eu

Niveau de maturité par discipline et par catégorie



Conclusions :

Peu de différences entre les disciplines

Niveau basique

Recommandations du « cookbook » : **sensibilisation**

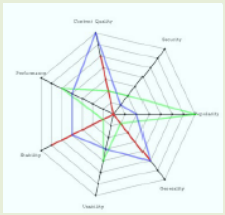
Actions de sensibilisation

SIM4RDM
SUPPORT INFRASTRUCTURE MODELS FOR RESEARCH DATA MANAGEMENT
http://www/sim4rdm.eu

« Institution »
OTELo

CHERCHEURS

Questionnaire



Niveau de maturité

« Infrastructure de soutien »
Inist-CNRS

Approches en présentiel

Bilan de l'étude

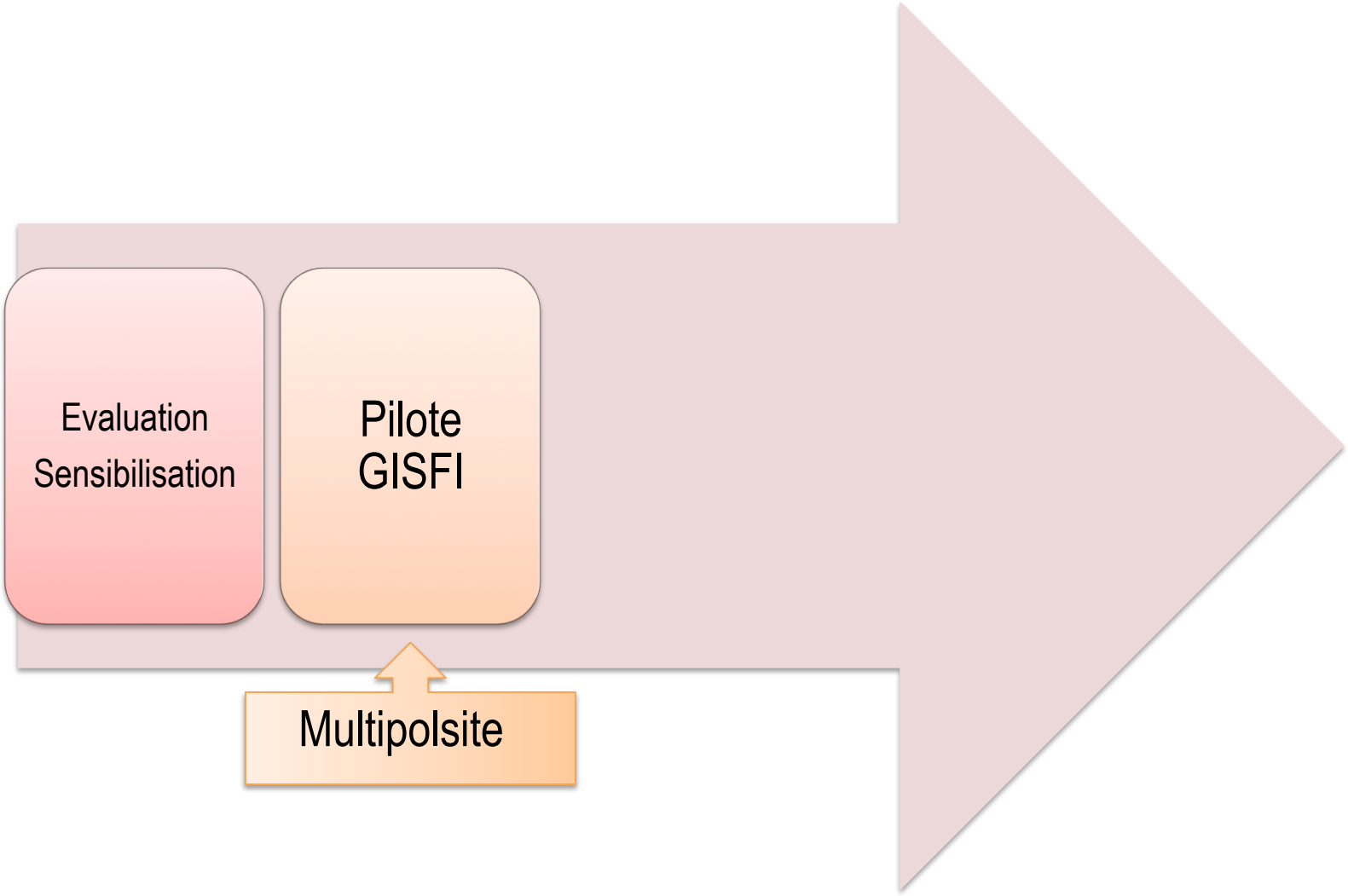
- Participation à un projet européen : dimension européenne, internationale
- Travail important de recherche et construction de contenu
- Mise en œuvre d'actions de sensibilisation et de conseil
- Expérience de communication
- Expérience pédagogique
- Curiosité des chercheurs éveillée

- Chercheurs peu disponibles
- Pas de politique des données nationale ou institutionnelle
- Pas d'infrastructure technique
- Chercheurs issus de communautés différentes
- Terminologie à adapter (« jargon »)
- Contenu trop riche en messages
- Pas suffisamment incitatif : montrer davantage les bénéfices pour le chercheur

Bilan de la sensibilisation :

	Chercheurs
Acquis	✓ Curiosité éveillée ✓ Acquisition des concepts de base, terminologie ✓ Réflexion autour des identifiants chercheurs ✓ Addition d'un paragraphe sur la gestion des données dans les propositions de projet ✓ Quelques « champions »
A développer	✓ Conviction des bénéfices d'une bonne gestion et partage des données (pour eux et puis pour les autres, visibilité, transparence, confiance, …) ✓ Appropriation de bonnes pratiques (mais pas de toutes les compétences IST!) ✓ Formation dès le Master + Bénéficier d'une reconnaissance

Stratégie d'OTELo



Des piliers

Objectifs

- Evolution vers une culture du partage
- Définition d'une politique des données
- Développement d'une infrastructure technique et humaine

Principes de base

- Engagement des chercheurs
- S'appuyer sur les usages pour faire évoluer les pratiques

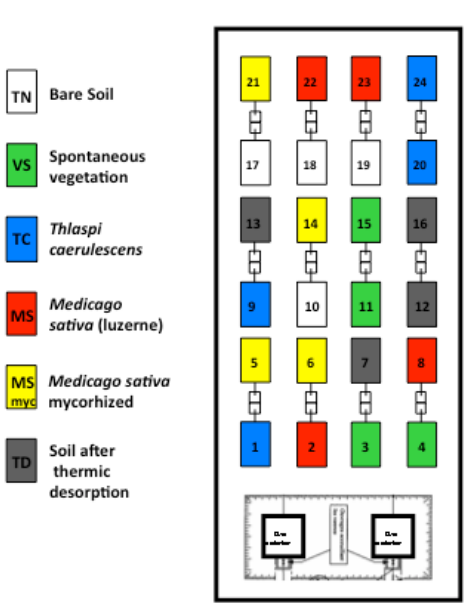
Acteurs

- **Chercheur/ingénieur** : contexte, **intelli**
- **Informaticien** : stockage, sauvegarde, **a**
- **Documentaliste scientifique** : documen
  standards, **interopérabilité**

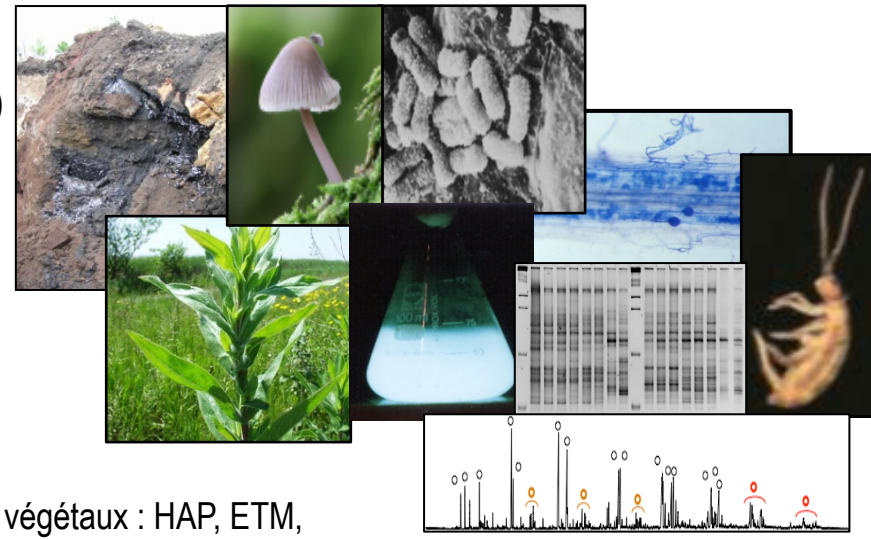
Reproductibilité
Réutilisabilité

Pilote : projet **MultipolSite** (ANR CESA 2008)

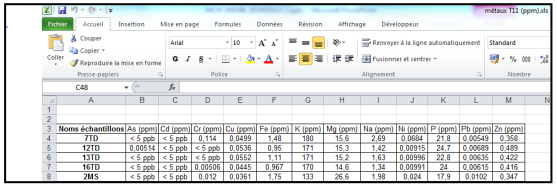
- Essais d'atténuation naturelle assistée par des plantes sur des terres de cokeries multipolluées
- Site expérimental du GISFI (Homécourt , 54)
- Suivi d'un réseau de parcelles et de lysimètres



Pilote : projet **MultipolSite** (ANR CESA 2008)



- **2 collectes d'échantillons/an de 2005 à 2013**:
  - Terres et percolats, végétaux (plantés ou spontanés)
- **Techniques et équipements :**
  - Observation: Loupes binoculaires, microscope
  - Analyse chimique: GC-MS, HPLC, ICP-AES,
  - Biologie moléculaire: PCR-TTGE, qPCR
  - Tests normalisés d'écotoxicologie
- **Paramètres mesurés :**
  - Nature des polluants dans les terres, les percolats et les végétaux : HAP, ETM,
  - Production de biomasse végétale
  - Structure des communautés, quantité et diversité des microorganismes (bactéries-champignons) et de la faune (macro et méso),
  - Évaluation de la toxicité du milieu



Noms échantillons	As (ppm)	Cd (ppm)	Cr (ppm)	Cu (ppm)	Fe (ppm)	K (ppm)	Mg (ppm)	Na (ppm)	Ni (ppm)	P (ppm)	Pb (ppm)	Zn (ppm)
7TD	< 5 ppb	< 5 ppb	0,114	0,0499	1,48	180	15,6	2,69	0,0684	21,8	0,00549	0,358
12TD	0,00514	< 5 ppb	< 5 ppb	0,0536	0,95	171	15,3	1,42	0,00915	24,7	0,00689	0,489
13TD	< 5 ppb	< 5 ppb	< 5 ppb	0,0552	1,11	171	15,2	1,63	0,00996	22,8	0,00635	0,422
16TD	< 5 ppb	< 5 ppb	0,00506	0,0445	0,967	170	14,6	1,34	0,00991	24	0,00615	0,416
2MS	< 5 ppb	< 5 ppb	0,012	0,0361	1,75	133	26,6	1,98	0,024	17,9	0,0102	0,347

→ *Stockage des données sur ordinateurs personnels (excel)*

- Mise en place d'une base de données **en cours de projet par un ingénieur en CDD** (stockage et suivi des données, analyse interdisciplinaire et intégration d'autres projets)
- **État de la base:**
  - Peu alimentée (problème de timing: base disponible en fin de projet, départ de l'ingénieur en CDD),
  - Ergonomie de la base

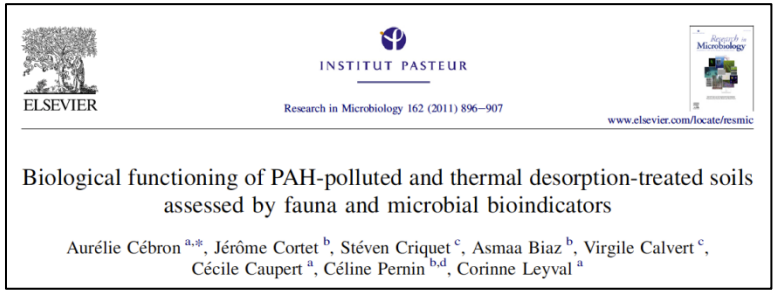
Apport du documentaliste : Bilan des pratiques de gestion

Description des données	Données d'observation et expérimentales (analyses génétiques, physicochimiques, …) Volumétrie faible : Mo Formats : .xls, .jpg, … Pas de convention de nommage des fichiers
Documentation **Standards des métadonnées**	Documentation dans le cahier de laboratoire ou dans les fichiers excel Nomenclature hétérogène Pas de standards de métadonnées appliqués (formats, taxonomie, thésaurus, vocabulaire contrôlé)
Partage des données	Pas de mise à disposition ou de diffusion
Stockage, sauvegarde **Archivage pérenne**	Ordinateurs personnels, intranet BDD créée pour le projet par un CDD : peu de données Pas d'archivage pérenne
Rôles et responsabilités	Mal définis

Réflexions côté chercheur

- Intégration dès l'écriture et mise en place de la base **dès le début du projet**
- La mise en place fait appel à des compétences que les chercheurs dans leur majorité ne possèdent pas :
  - Informaticiens
  - Documentalistes scientifiques
- **Vision collective**

- Prise de conscience de l'importance de la base: il faut que le chercheur y trouve une **plus-value** :
  - qualité des données stockées et informations associées
  - analyse multidisciplinaire des données
  - Réutilisation ultérieure


ELSEVIER

INSTITUT PASTEUR

Research in Microbiology 162 (2011) 896–907

www.elsevier.com/locate/resmic

Biological functioning of PAH-polluted and thermal desorption-treated soils assessed by fauna and microbial bioindicators

Aurélie Cébron [a,*], Jérôme Cortet [b], Stéven Criquet [c], Asmaa Biaz [b], Virgile Calvert [c], Cécile Caupert [a], Céline Pernin [b,d], Corinne Leyval [a]

- **Besoin de planification, organisation de la gestion des données**

- **Valorisation** possible de la base : publication…
  mais avec les craintes quant à la perte du contrôle des données

Elaboration d'un exemple de plan de gestion Multipolsite

1.1.1.1 Diversité mycorhizienne :

Méthode : la diversité des champignons mycorhiziens à arbuscules des parcelles à une date donnée est analysée par le biais d'une empreinte moléculaire. La méthode employée n'est pas une méthode normalisée. L'extraction d'ADN est conduite sur un échantillon moyen de racines par parcelle. Une seule PCR nichée est réalisée pour chaque échantillon d'ADN.

*SONJAK S, **BEGUIRISTAIN T.**, LEYVAL C. AND REGVAR M., 2009 Temporal temperature gradient gel electrophoresis (TTGE) analysis of arbuscular mycorrhizal fungi associated with selected plants from saline and metal polluted environments. **Plant and Soil**, 314 (1-2) 25-34.*

Stockage des échantillons d'ADN : Les échantillons d'ADN concentrés et dilués ont été conservés dans un congélateur à -20°C. (LIEC AIguillette)

Données sélectionnées pour conservation :

Données	Formats	Stockage	Durée de conservation
Photos des gels de la matrice	.jpg	Serveur	> 10 ans
Matrice	.xls (Excel)	Serveur	> 10 ans
Diagramme MDS	.jpg	Serveur	> 10 ans
Métadonnées	MySQL	Base de données	

Photos de gels de la matrice sur Serveur :
http://www.gisfi.fr/documents/diversite_mycorhizienne/matrices-tous-les-temps.xlsx
Métadonnées dans la Base de données (Mysql) : http://www.gisfi.fr/gisfi-SystemeInfo/consulter.php

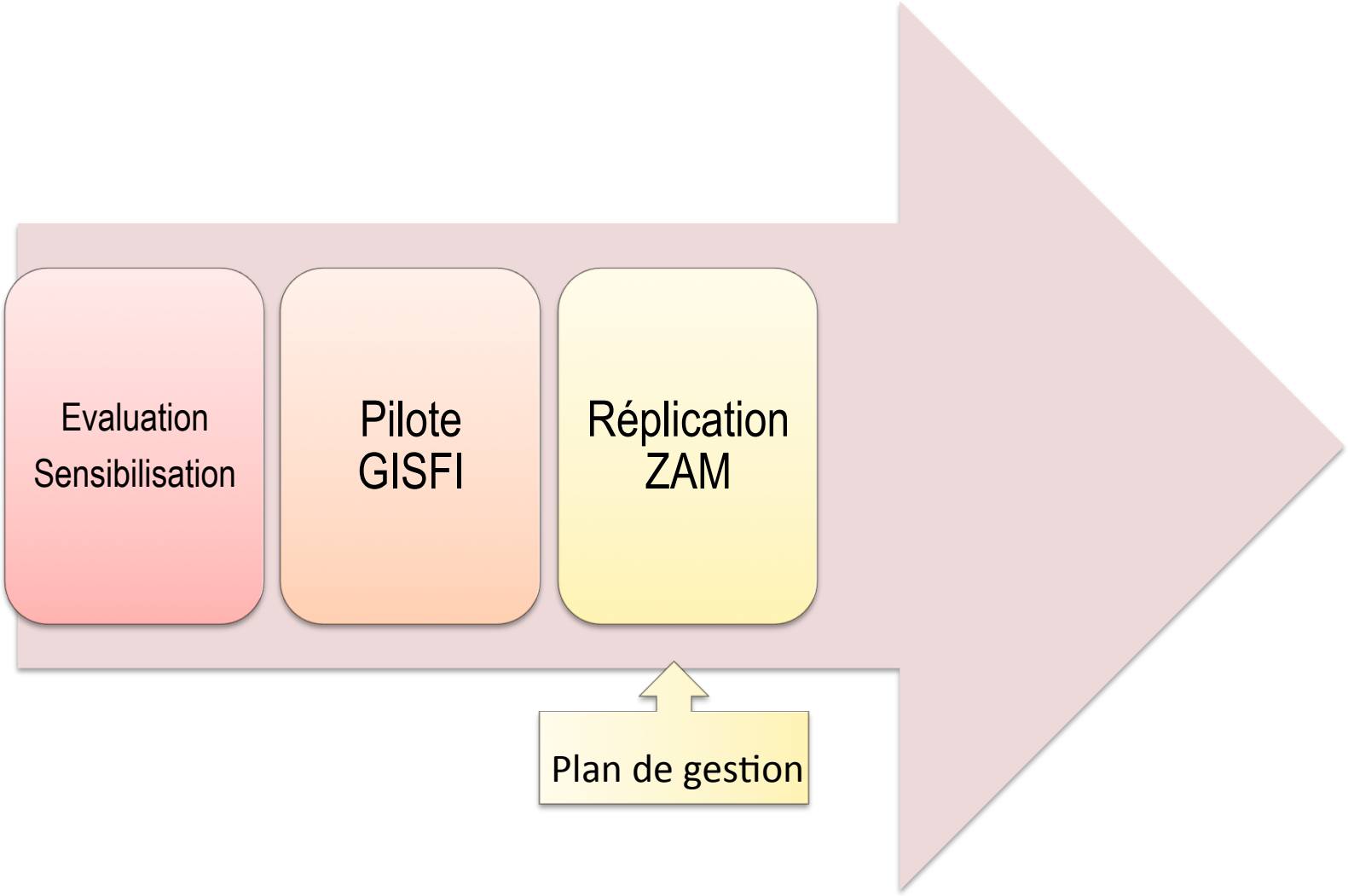
Bilan d'étape de la phase pilote

Observations/réflexions

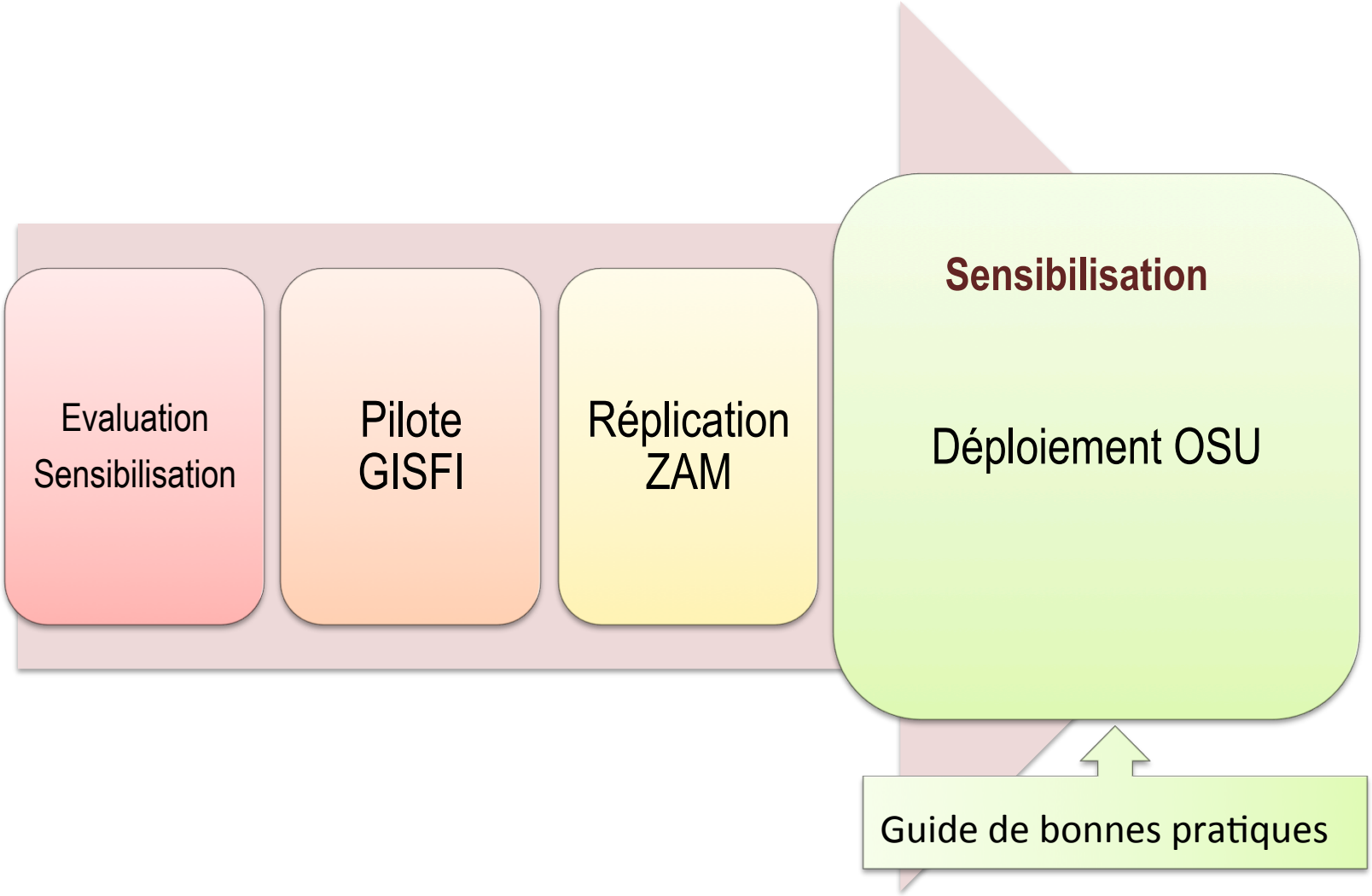
- Appréhension du rôle d'accompagnement des chercheurs
- S'approprier la thématique scientifique
- Utiliser l'existant pour améliorer les pratiques et entraîner les chercheurs dans cet effort de gestion
- Ecouter/comprendre les besoins et réticences des chercheurs (« empathie »)
- Adaptabilité, souplesse, curiosité
- Poser des questions pour susciter la réflexion
- Patience, persévérance : temps de maturation, d'appropriation par les chercheurs
- Importance de la complémentarité: chercheurs - informaticiens - documentalistes

->Engagement des chercheurs : réflexion sur le plan de gestion des données et la conservation des données

Stratégie d'OTELo



Stratégie d'OTELo



Elaboration d'un guide de bonnes pratiques

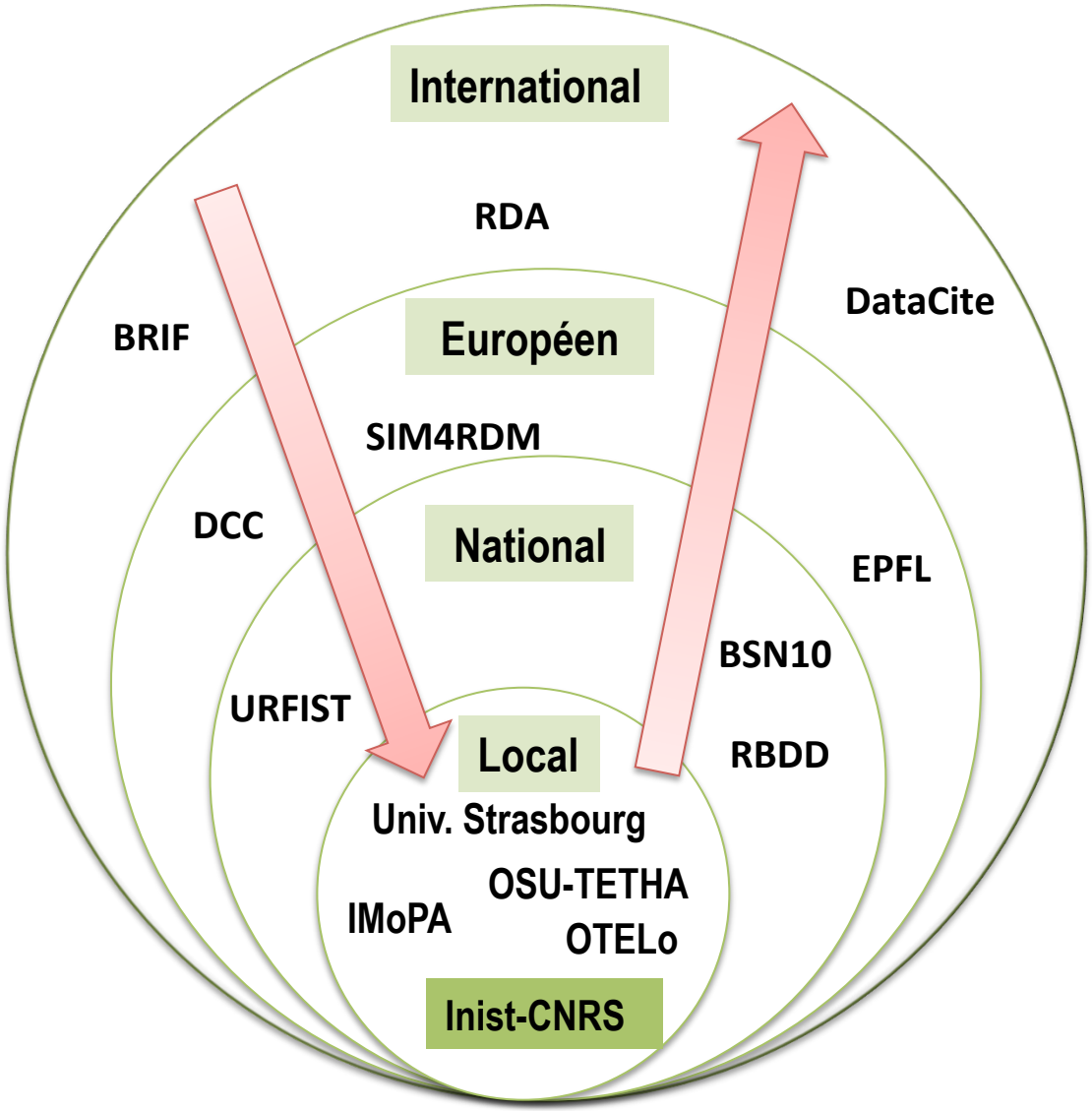
- **Auteurs** : chercheurs, ingénieurs, informaticien, documentalistes
- **Objectifs** :
  - Recommandations, des conseils et des outils
  - Document synthétique, exemples
  - Pédagogique : sens à la gestion des données, pas hermétique
- **Contenu**
  - Conventions de nommage, organisation des fichiers
  - Structuration des données, dictionnaire des données
  - Documentation des données : template pour les métadonnées
  - Valorisation: visibilité, réutilisabilité

Evolution de la réflexion en interne

Mobised ZAM

CopliHo
(GISFI)

Des rôles auprès des chercheurs ?



Récepteur
Transmetteur

Coordinateur

Facilitateur

Collaborateur

Pédagogue

Take-home messages

10 key tips from practitioners on how to get started with research data management

1. Start with the facts (not with 'how things should be')
2. Teach each other
3. Don't wait; get going
4. Focus on opportunities
5. Be fearless
6. Gather good practices
7. Have an interest in researchers and research practice
8. Offer something to make a researcher's day easier
9. Keep on fine-tuning and reinventing
10. Keep on research dating

Conclusions déduites à partir de rapports de 11 études de cas et du workshop organisé par le comité de pilotage "**Scholarly Communication and Research Infrastructures " LIBER 43rd annual conference (2014)**

S'il n'y a que des contraintes,
les chercheurs n'auront pas le « spirit of the law »
(Borgman CL, RDA 4th Plenary Meeting, 2014)

Donnez du sens à la gestion des données
Mettez en exergue les bénéfices

Remerciements

Direction d'OTELo

Corinne Leyval

Frédéric Villiéras

Pierre-Yves Arnould, informaticien et porteur du projet de gestion des données

Membres d'OTELo

Léa et Eléonore Jacquemot