



www.cnrs.fr

Visa pour le Text-Mining

Stéphane Schneider
Fabienne Kettani

INIST-CNRS



TDM : Le problème

⊙ Masse d'information croissante

- La recherche mondiale génère **2,5 millions** d'articles par an
- Un article publié toutes les **12 secondes**,
- **70000** articles rédigés uniquement sur le suppresseur de tumeur p53,



⊙ Aujourd'hui, les techniques de TDM permettent à la machine de mieux comprendre et prédire

TDM : Paysage complexe

TYPE DE TEXTE

Littérature scientifique,
Tweets / blogs,
Brevets,
Registres cliniques,
Manuels scolaires,
Forums en ligne

DOMAINES

Santé,
Biologie,
Sciences sociales,
Humanités,
Finances/affaires,

TACHES

Extraction Information,
Recherche
Sémantique,
Question/Réponse,
Analyse de Sentiment,
...

METHODES

Machine learning,
Recherche Pattern
Sac de mot,
Clustering,
Dictionary lookup,
...

ACTEURS

Chercheur TDM
Intégrateur TDM
Fournisseur de contenu
Chercheur Users
Infrastructure Data



LANGUES

Anglais,
Français,
Allemand,
Espagnol,
Portugais,
Italien,

NORMES, FORMATS

LEMON, SKOS,
OLIF, UIMA , TEI ,
LMF, OBO, OWL,
ONTOLEX,XLIFF,
LFG, TAG...
...

TDM : Les besoins



- Besoins grandissants de traitements TDM spécifiques
- Capitalisation et partage des solutions TDM
- Utilisation par des non-spécialistes du TDM
- Centralisation de l'expertise de développement des applications
- Moyen de calcul pour des traitements à grande échelle
- Associer bibliothèques numériques et communauté des chercheurs en TDM

TDM : Comment faire ?



Fournir un cadre d'interopérabilité pour le traitement TDM



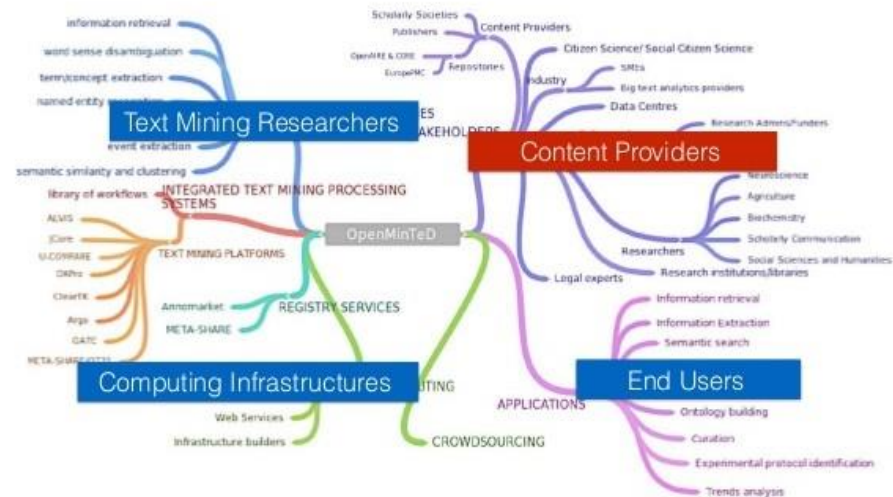
Développer une plateforme orientée service



Partager des contenus : input et résultats



Dresser un pont entre toutes les communautés concernées



OpenMinted : Organisation



European Commission



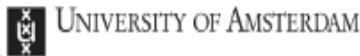
Projet d'infrastructure européenne de text-mining Open Source

Durée : 1er juin 2015 à fin en mai 2018

16 partenaires - Coordination de l'Université d'Athènes

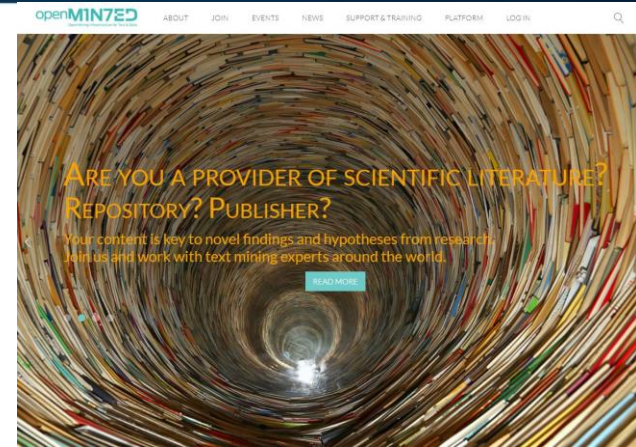
Les équipes européennes expertes en conception de plateforme de TDM ont développées une plateforme :

- Usine à fabriquer des offres TDM : la composition, réutilisation, exécution de workflows
- Bibliothèque riche de composants TDM interopérables
- Connection aux infrastructures existantes dont bibliothèques numériques et portails de ressources sémantiques (ontologie/termino)
- De conception « modulaire » et évolutive
- Assure la sécurisation juridique des pratiques



OpenMinted : Objectifs

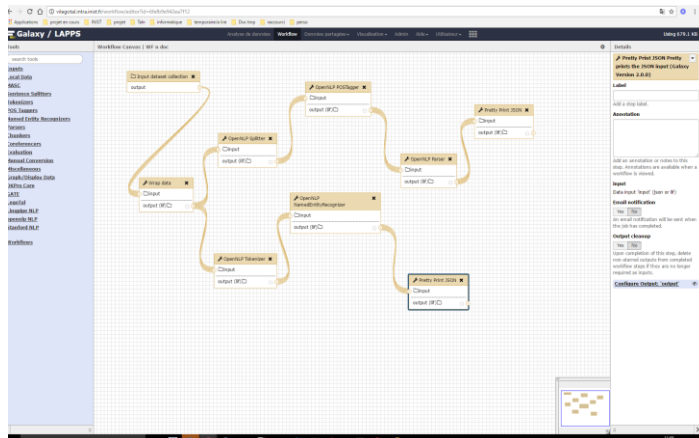
Proposer une infrastructure de Text et Data Mining (TDM), **ouverte et pérenne**, qui permette aux chercheurs un accès facilité aux technologies text-mining applicables à un vaste éventail de sources de la littérature scientifique.



www.cnrs.fr

open
**MIN
TED**

<http://openminted.eu/>



Permettre aux acteurs TDM de **partager leurs outils, leur corpus, ou des résultats** ainsi que de créer et de partager des workflow d'analyse.

OpenMinted : La plateforme

APPLICATIONS UTILISATRICES TDM

CHERCHEURS



EXPERTS TDM



S
E
R
V
I
C
E
S

CONCEPTEURS TDM

TOOLS



WORKFLOW



CORPUS

RESULTS

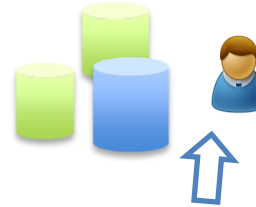
REGISTRE + DATA



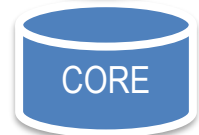
FOURNISSEURS TDM



UTILISATEURS
RESULTATS



FOURNISSEURS DE
CONTENU



VisaTM : Objectif

va tenter d'apporter des éléments pour **CONVAINCRE** de l'**OPPORTUNITE** pour un **OPENMINTED** France

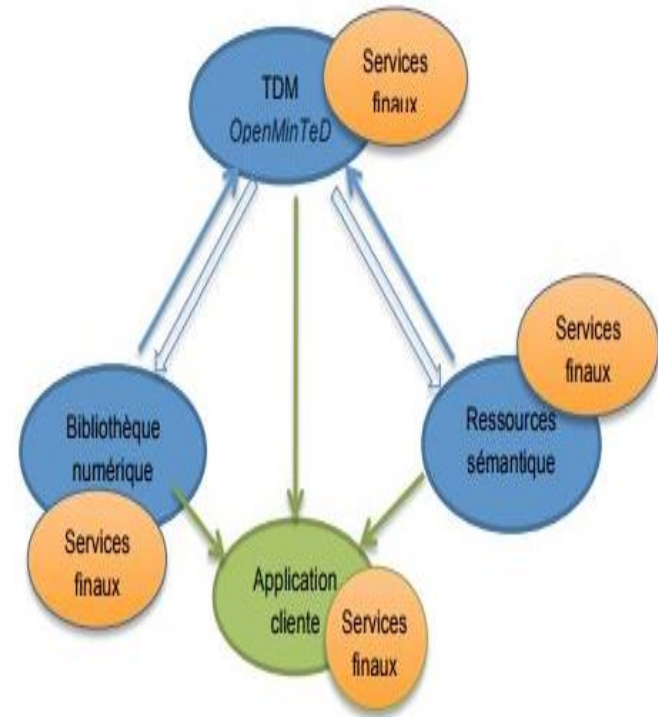


www.cnrs.fr

L'**objectif** du projet est d'étudier les conditions de production de services TDM à haute valeur ajoutée basés sur l'analyse sémantique, pour les chercheurs **en France**.

qui repose sur **trois infrastructures** de type

- **ISTEX** comme bibliothèque numérique
- **OpenMinted** comme plateforme TDM
- **Agroportal** comme portail de ressources sémantiques



VisaTM : Organisation



Partenariat

Multi-établissements CNRS, Université Montpellier et établissements de recherche finalisés (Inra Bibliome, Dist Inra, ...).

Pilotage

CoSO (BSN)

Budget

160 000 euros

Durée

18 mois

VisaTM : Missions dans la future plateforme TDM

Création de contenu

- ✓ Conception/adaptation/réutilisation de briques logicielles
- ✓ Création et mise à disposition de ressources (termino...)
- ✓ Mise à disposition de corpus et de résultats



www.cnrs.fr

Service d'aide, support

- ✓ Aide à l'intégration de brique logiciel
- ✓ Aide à la connexion de fournisseur de contenu
- ✓ Aide à l'intégration de services (aide à faire des applications utilisant des services OMTD)

Conseils, animation

- ✓ Etudes de cas et définition de guides de bonne pratiques
- ✓ Conseil sur les aspects légaux autour du TDM (licence, copyright..)
- ✓ Animation de la communauté
- ✓ Promotion de la plateforme

Administration de la plateforme

- ✓ Contrôle qualité sur les dépôts
- ✓ Gestion technique de l'hébergement de la plateforme

Visa™ : 1/3 chantiers

Connecter techniquement ISTEEX et AGROPOTAL à OMTD

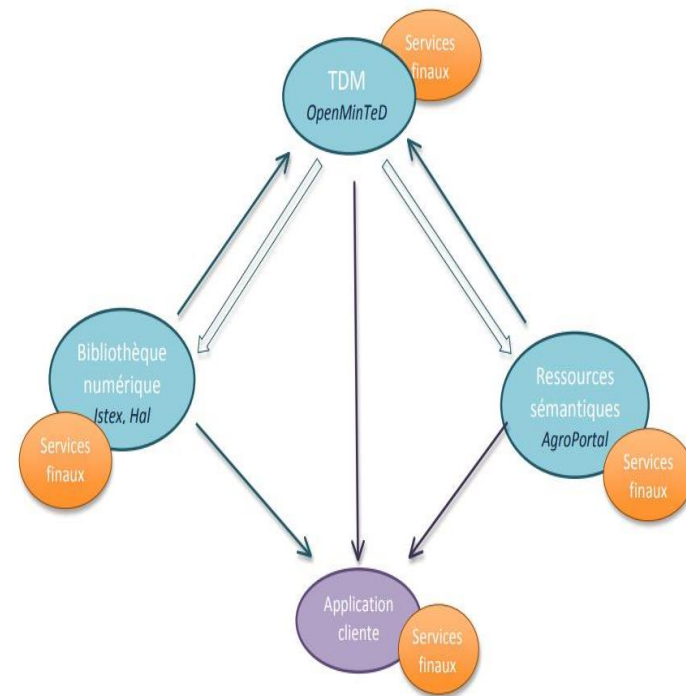


www.cnrs.fr

Pour créer une synergie française entre infrastructures :

- **ISTEX** comme bibliothèque numérique
- **OpenMinted** comme plateforme TDM
- **Agroportal** comme portail de ressources sémantiques
- **les infrastructures métiers** pour répondre au plus près du besoin du chercheur

VISA™, une synergies entre 3 piliers



Visa™ : 2/3 chantiers

Développer deux applications pilotes qui utilisent ces connexions



INRA : Dans une application bioinformatique, intégrer des résultats d'extraction d'information de textes scientifiques à des données d'observation et expérimentales

Usage : Découverte de nouvelles connaissances




INIST : Sur un corpus ISTEEX dans le domaine des Géosciences, un utilisateur applique à la volée des traitements TDM permettant d'avoir un aperçu thématique des documents

Usages : Constitution et exploration de corpus (délimitation)

VisaTM : 3/3 chantiers: Objectifs

Réalisation d'une grande étude pour

- 
- www.cnrs.fr
- ① Etudier les conditions de production de services de TDM à haute valeur ajoutée pour les chercheurs
 - Panorama du contexte
 - Scénarios
 - ② Imaginer et décrire une infrastructure technique et humaine nationale dont une instance OMTD constituerait un nœud pour:
 - Interconnecter des infrastructures nationales et internationales de fourniture de documents (bibliothèques numériques) et/ou de ressources sémantiques
 - Fournir un service de TDM
 - Aux chercheurs français (spécialistes en TAL/TDM et autres disciplines)
 - Aux décideurs
 - ③ Etablir une feuille de route de la mise en place du dispositif

VisaTM : 3/3 chantiers : Livrables

🕒 LIVRABLE 1: Panorama du contexte

○ Cartographie des acteurs et des compétences nécessaires

- Producteurs/fournisseurs de contenus/d'outils
- Utilisateurs: chercheurs , décideurs, intégrateurs

○ Cartographie des outils

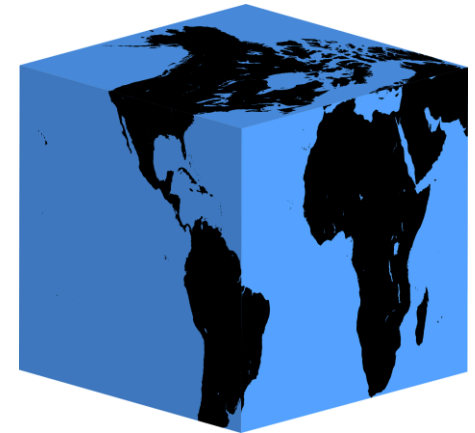
- Plateformes : OMTD
- Outils libres
- Outils payants

○ Recueil des besoins

- Suivant la typologie des acteurs
- Indicateurs du succès: utilité, utilisabilité, performance, pérennité, coût de réadaptation

○ Analyse de l'environnement

Aspects politiques, économiques, scientifiques, techniques, juridiques (Directive européenne sur le droit d'auteur et Loi numérique)



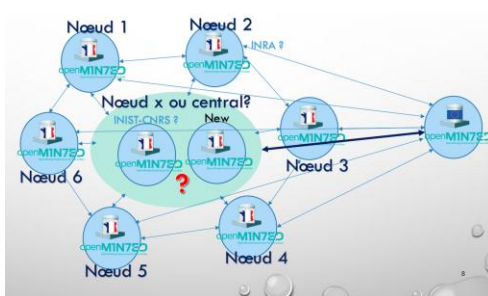
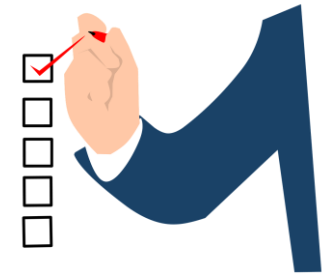
VisaTM : 3/3 chantiers : Livrables

🕒 LIVRABLE 2: Scénarios

Soumission à un comité d'experts afin de dégager des scénarios à privilégier.



- Une base d'invariants
 - Choix de l'hébergeur d'OMTD France par exemple
 - ...
- Des scénarios suivant les communautés d'utilisateurs
- Des scénarios en fonction de choix techniques/de fonctionnement



- Centralisation sur une structure existante ou création d'une structure dédiée
- Réseau de nœuds OMTD avec un nœud central et des satellites dont chacun possède toutes les fonctionnalités et suivant un déploiement local ou non
- Réseau de nœuds OMTD avec un nœud central et des satellites et des fonctionnalités réparties

...

VisaTM : 3/3 chantiers : Livrables

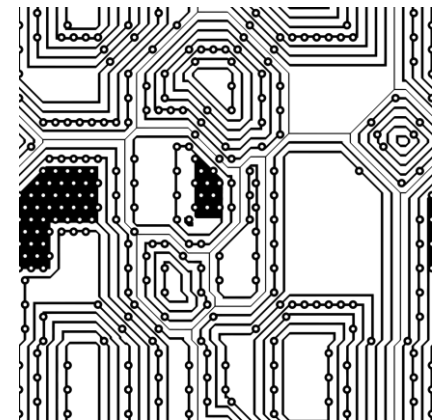
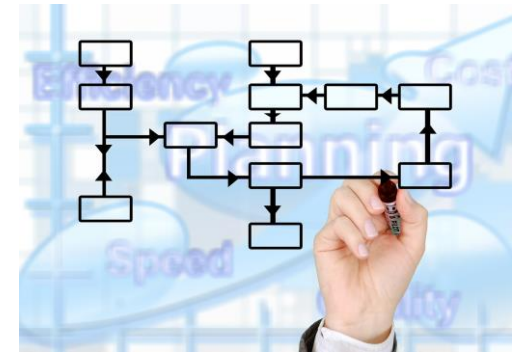
🕒 LIVRABLE 3: Description du/des dispositif(s) visé(s)

○ Aspects organisationnels

- Gouvernance et coordination
- Financement
- Compétences/formation nécessaires
- Modèle d'exploitation
 - ✓ Description de l'infrastructure: outils, ressources, accès,...
 - ✓ Description de l'offre de service: accompagnement, animation de communautés, besoins en outils interconnectés, en interfaces utilisateurs,...
 - ✓ Evolution: montée en charge,...
- Juridique

○ Aspects techniques

- Infrastructure
 - ✓ Matérielle/logicielle
 - ✓ Calibrage des besoins
 - ✓ Ressources RH
 - ✓ Déploiement
- Appui sur des initiatives proches et les choix faits pour OMTD



VisaTM : 3/3 chantiers : Livrables

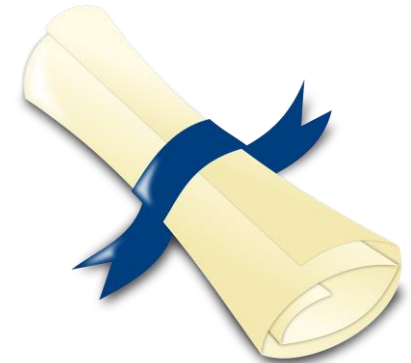
🕒 LIVRABLE 4: Feuille de route et calendrier

Description de la mise en œuvre du dispositif et différentes étapes dans le temps.



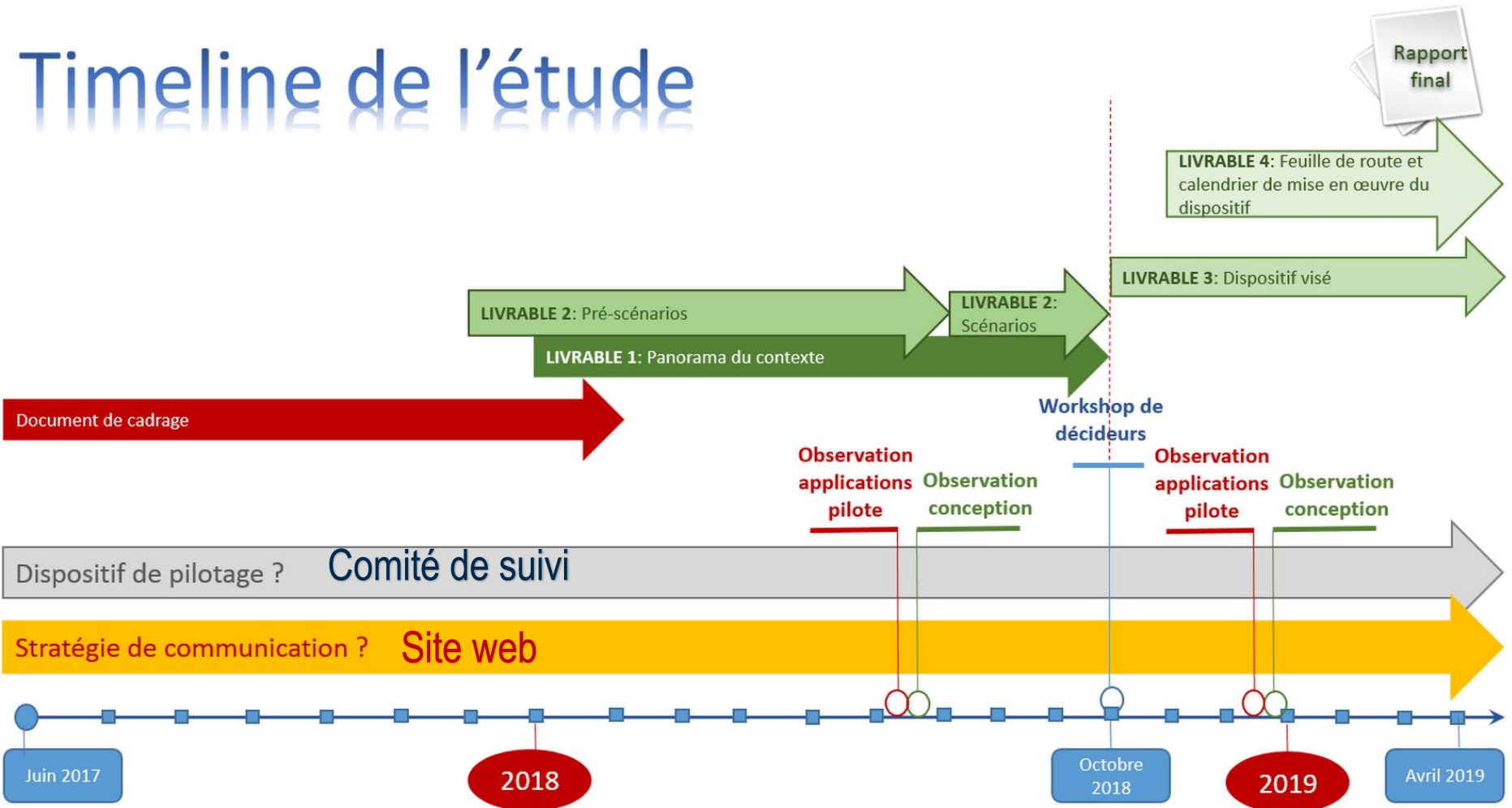
🕒 RAPPORT FINAL

- Analyse critique documentée et argumentée
- Présentation des points forts du projet
- Recommandations



VisaTM : 3/3 chantiers : déroulement

Timeline de l'étude





www.cnrs.fr

Merci !

