

Regroupements et alignements de données internes à la BnF

Des outils, des données et des
humains

Sébastien Peyrard, BnF
sebastien.peyrard AT bnf.fr

Adrien Di Mascio, Logilab
adrien.dimascio AT logilab.fr

Plan

- La démarche : exploiter des algorithmes pour améliorer les données de la BnF
- La situation actuelle : des outils de traitement mis à disposition des experts
- Perspectives

La démarche

- Objectif : améliorer les données de la BnF :
 - Personnes, œuvres
 - Les éditions ne sont pas toujours rassemblées sous un même chapeau œuvre
 - Exemples :
 - Notice BIB liée à son Œuvre
 - Notice BIB non liée à une œuvre alors que l'œuvre existe → alignement
 - Notice BIB non liée, l'œuvre n'existe pas → clustering
- ➔ Générer des œuvres partout

Un exemple : un auteur pour lequel on a des œuvres...

Michel Houellebecq : *pseudonyme individuel*



Pays : **France**

Langue : **français**

Sexe : masculin

Note : Scénariste, poète, essayiste. - Le prix Goncourt 2010 a été attribué à Michel Houellebecq pour le roman " la carte et le territoire". - Ingénieur agronome. - Photographe

Domaines : Littératures

ISNI : **ISNI 0000 0001 2137 320X**

Ses activités
(115 documents)

Documents à propos
de cet auteur

Pages dans data.bnf.fr
(2 pages)

Sources
et références

→ **Extension du domaine de la lutte (1994)**

Roman

→ **Les particules élémentaires (1998)**

Roman

→ **Plateforme (2001)**

Roman

→ **Soumission (2015)**

Roman

→ [La carte et le territoire](#)
Description matérielle : 1 vol. (428 p.)
Édition : Paris : Éd. de Noyelles , impr. 2010
Auteur du texte : [Michel Houellebecq](#)
[\[catalogue\]](#)

→ [La carte et le territoire](#)
Description matérielle : 1 vol. (428 p.)
Édition : [Paris] : le Grand livre du mois , impr. 2010
Auteur du texte : [Michel Houellebecq](#)
[\[catalogue\]](#)

→ [La carte et le territoire](#)
Description matérielle : 1 vol. (487 p.)
Description : Note : Bibliogr. p. 483-487
Édition : [Paris] : Flammarion , DL 2016
Auteur du texte : [Michel Houellebecq](#)
Éditeur scientifique : [Agathe Novak-Lechevalier](#)
[\[catalogue\]](#)

→ [La carte et le territoire](#)
[Éd. en gros caractères]
Description matérielle : 1 vol. (511 p.)
Édition : Cergy : À vue d'oeil , impr. 2011
Auteur du texte : [Michel Houellebecq](#)
[\[catalogue\]](#)

→ [La carte et le territoire](#)
Description matérielle : 1 vol. (431 p.)
Édition : Paris : Éd. France loisirs , impr. 2011
Auteur du texte : [Michel Houellebecq](#)
[\[catalogue\]](#)

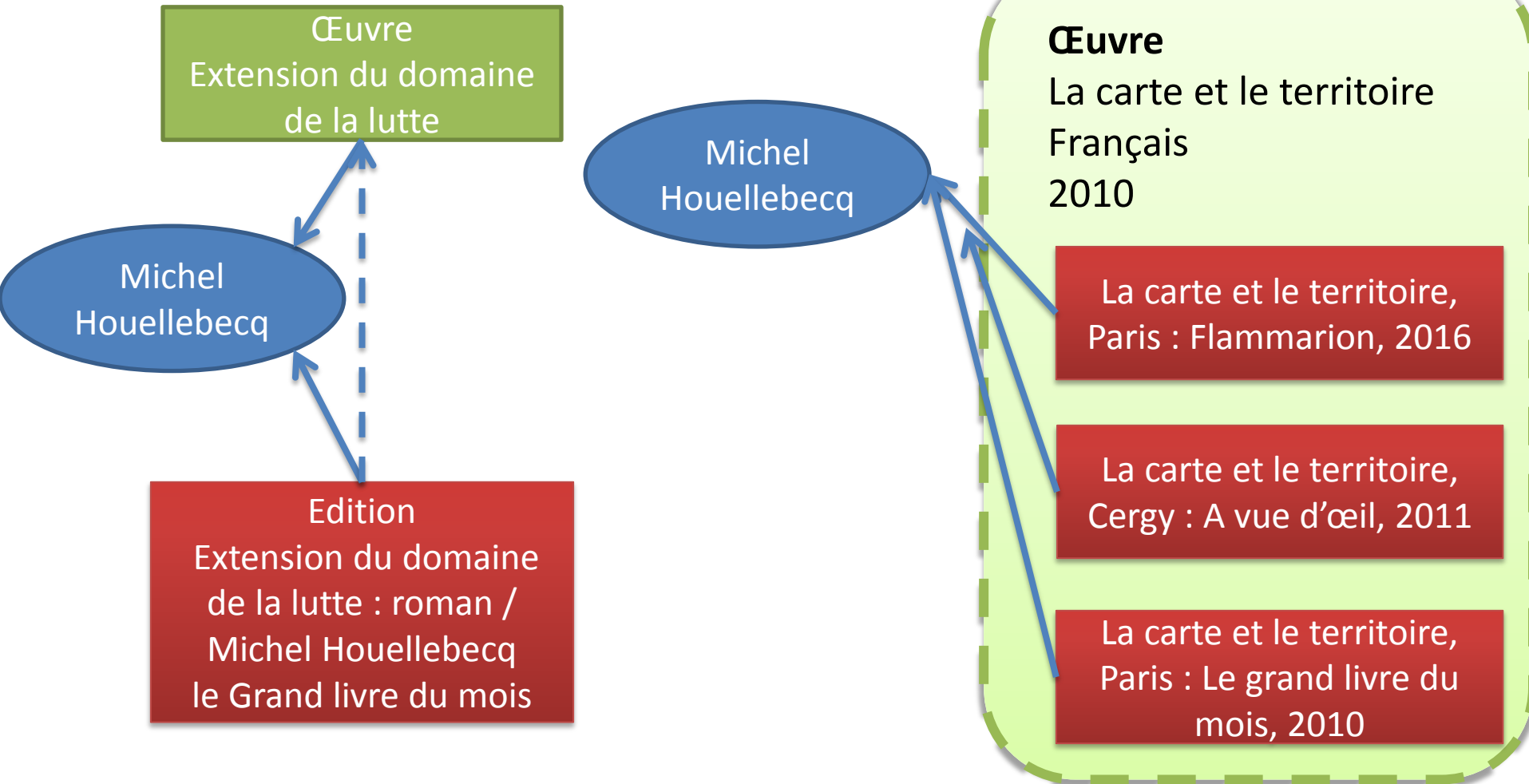
→ [La carte et le territoire](#)
roman
Description matérielle : 1 vol. (413 p.)
Édition : Paris : J'ai lu , impr. 2012
Auteur du texte : [Michel Houellebecq](#)

... ou pas
(majorité des cas)

La situation actuelle

Aligner des éditions à des œuvres

Créer des œuvres



Perspectives et interrogations

- Cycle de vie des données :
 - versement des œuvres dans data.bnf.fr
 - versement des œuvres dans le catalogue général à partir du moment où les informations sont stabilisées
- Enrichissement itératif des données d'œuvre : éviter de prévoir 3 coups à l'avance, s'appuyer sur les premiers calculs pour déduire de nouvelles informations
- Modulariser la génération des métadonnées de niveau œuvre (actuellement prévu pour les œuvres textuelles du 20^e siècle)
- Sur quelles fonctionnalités donner la main à l'expert ? Trouver la bonne frontière
- Un outil qui pourrait être élargi à d'autres bases sources et d'autres traitements

Conclusion

- Un outil qui permet de gagner en autonomie sur la manipulation de ses propres données
- L'expertise sur les données est déjà là, donnons-lui les moyens d'être plus efficace