abes
agence bibliographique
de l'enseignement supérieur

theses.fr

Guide de réutilisation des données de theses.fr

Mise à jour mai 2022

Ensemble, avec nos réseaux,
réinventons le service public des données

abes.fr

Ce guide est destiné aux utilisateurs des données et métadonnées de theses.fr. Il présente de manière synthétique les principales caractéristiques (périmètre, source, exhaustivité) ainsi que la structuration (format, modèle, référentiels) des jeux de données disponibles.

Table des matières

Données disponibles à partir de theses.fr

Complétude des données

Conformément à l'arrêté du 25 septembre 1985, l'arrêté du 7 août 2006 et l'arrêté du 25 mai 2016, le signalement des thèses de doctorat françaises soutenues depuis 1985 constitue une obligation réglementaire.[1]

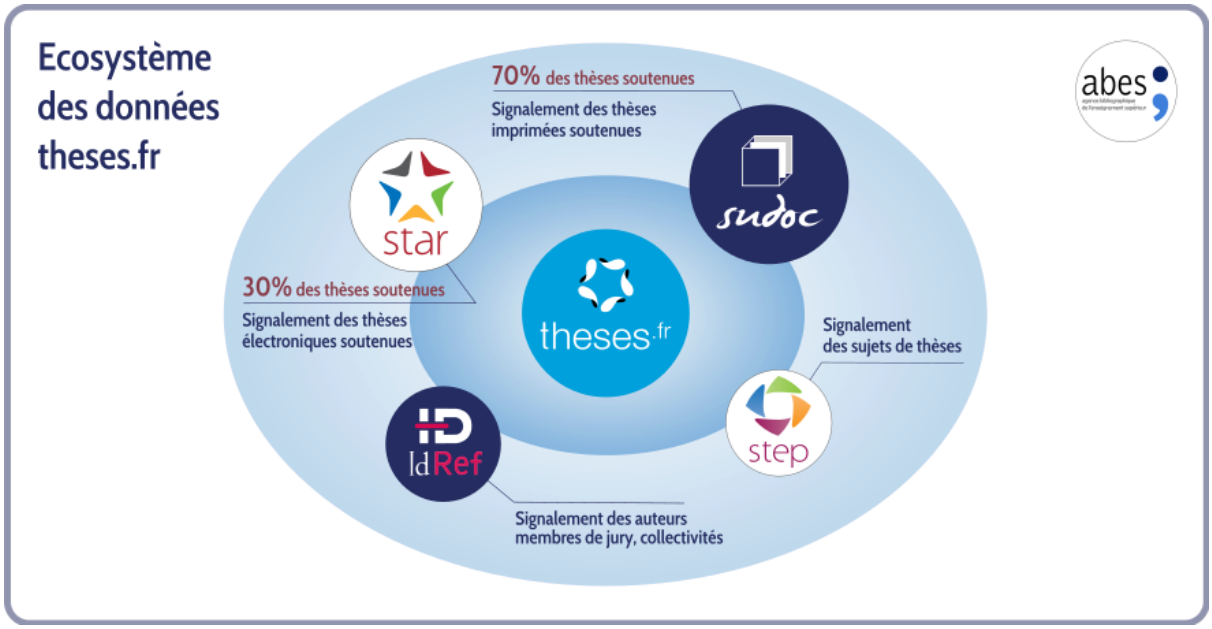
Ainsi, le moteur de recherche theses.fr recense de manière quasi exhaustive les thèses de doctorat soutenues en France depuis 1985 ainsi que les sujets de thèses en préparation.

Les données disponibles sont produites par les établissements habilités à délivrer le doctorat et sont placées sous leur responsabilité. Selon les périodes et les établissements, les données concernant ces thèses peuvent présenter des lacunes. Certaines informations, comme la liste des membres du jury ou les liens vers le référentiel IdRef, sont parfois absentes. **Les données des deux dernières années sont non consolidées et doivent être considérées comme non exhaustives**. Les délais de traitements des thèses, propres à chaque établissement, influent sur la complétude de ces données. Pour l'année 2021, le délai de traitement moyen entre la date de soutenance de la thèse et sa publication effective dans theses.fr était de 250 jours.

Origine des données (au 1er février 2022)

Trois sources de données alimentent theses.fr :

- le Sudoc : métadonnées des thèses produites et archivées au format imprimé entre 1985 et 2016
- STAR : métadonnées des thèses produites et archivées au format électronique à partir de 2006
- STEP : métadonnées des sujets de thèse en cours de préparation[2]



[1] Les HDR, les thèses d'exercice et les mémoires ne font pas partie du périmètre de theses.fr.

[2] Ouverte en 2011, la base de données STEP intègre le Fichier Central des Thèses, créé et maintenu par Paris Nanterre entre 1968 et 2009, pour signaler les sujets de thèses en LSHS, ainsi que la base de données THESA, qui recensait les sujets de thèses en sciences de l'ingénieur.

Source / Années de soutenance / ou de préparation	1970-1985	1985-2006	2006-2016[3]	2016-….
Sudoc	✗	✓	✓	✗
STAR	✗	✗	✓	✓
STEP	✓	✓	✓	✓

Tableau 1: Provenance des données de thèses disponibles à partir de theses.fr

Thèses signalées dans le Sudoc

- **Période concernée :** 1985-2016
- **Type de document :** thèses produites et archivées au format imprimé
- **Volumétrie :** 70 % des thèses soutenues référencées sur theses.fr proviennent du Sudoc
- **Exhaustivité :** les données Sudoc font l'objet d'une sélection préalable. Seules les données de qualité suffisante sont versées dans theses.fr[4]

Thèses signalées dans STAR

- **Période concernée :** 2006-….
- **Type de document :** thèses produites et archivées au format électronique
- **Volumétrie :** 30 % des thèses soutenues référencées sur theses.fr proviennent de STAR
- **Exhaustivité :** l'utilisation de STAR constituant une obligation réglementaire, à quelques rares exceptions (absence de dépôt par le doctorant, thèse en cours de traitement par l'établissement de soutenance), les données de STAR sont exhaustives.

Sujets de Thèses signalés dans STEP

- **Période concernée :** 1970-….
- **Type de document :** sujets de thèses en préparation.
- **Exhaustivité :** l'utilisation de STEP ne constituant pas une obligation réglementaire, certains établissements habilités à délivrer le doctorat ne signalent pas les sujets de thèses en préparation. Les doctorants peuvent également s'opposer à la diffusion de leur sujet de thèse. Lorsqu'une thèse est signalée comme soutenue dans STEP, les données sont versées automatiquement dans l'application STAR. Une fois la thèse traitée et archivée, les données STAR viennent écraser les données STEP.

3 Entre 2006 et 2016, deux modalités de dépôt des theses étaient proposées aux établissements : le dépôt papier (traitement via l'application WinIBW) et le dépôt électronique (traitement via l'application STAR).

4 Liste des contrôles qualité réalisés avant import dans theses.fr

Récupérer les données de theses.fr

Les données peuvent être récupérées aux formats JSON, NDJSON ou CSV :

- à partir de l'API de thèses.fr
- via un dump actualisé tous les ans et mis à disposition sur data.gouv.fr.

Caractéristiques du jeu de données

Thèses soutenues en France depuis 1985

> *Ce jeu de données est constitué de l'ensemble des métadonnées descriptives des thèses de doctorat soutenues en France depuis 1985. Les thèses en préparation ne font pas partie des données…*

Récupérer le jeu de données data.gouv.fr

Couverture temporelle	**1985 - ….**
Couverture spatiale	France
Mise à jour	Annuelle
Licence	Licence Ouverte V 2.0
Formats	JSON, NDJSON, CSV
Fichiers	https://www.data.gouv.fr/fr/datasets/r/d4f0a317-4fd7-4850-bfa2-829a2a4a21df (JSON) https://www.data.gouv.fr/fr/datasets/r/78d463e4-b787-4b1f-86d5-e52e4fb86f1d (NDJSON) https://www.data.gouv.fr/fr/datasets/r/041e3d9a-6ada-41ca-af85-8e0d3838f605 (CSV)
Encodage	UTF-8
Séparateur de colonne (CSV)	,
Séparateur de chaîne de caractères (CSV)	"

Tableau 2, jeu de données disponibles : "Thèses soutenues en France depuis 1985"

Exposition et réutilisation des données personnelles

La mise à disposition de données relatives aux personnes se fait en conformité avec l'article D.312-1-3, paragraphe 4, du Code des relations entre le public et l'administration et l'article 17, paragraphe 3, du Règlement Général sur la Protection des Données. Les données publiées sont toutes des données publiques. Elles sont réduites au strict nécessaire, conformément à l'article 89, paragraphe 1, du Règlement Général sur la Protection des Données : noms et prénoms des personnes liées aux thèses de doctorat (auteurs, directeurs de thèse, membres du jury), liste des thèses liées au nom desdites personnes.

Les personnes concernées disposent d'un droit d'accès et de rectification des données, qu'elles peuvent exercer via le guichet d'assistance de l'Abes. Le droit à l'effacement ne peut s'exercer concernant le signalement des thèses de doctorat soutenues, conformément aux exceptions prévues par l'article 17, paragraphes 3.b et 3.d, du Règlement Général sur la Protection des Données.

Structuration des données :

- **accessible :** accessibilité de la thèse sur theses.fr (oui, non)
- **auteurs.[n].idref :** identifiant IdRef de l'auteur
- **auteurs.[n].nom :** nom de l'auteur
- **auteurs.[n].prenom :** prénom de l'auteur
- **cas :** scénario de diffusion de la thèse
- **code_etab :** code court de l'établissement de soutenance. Le code court identifie les environnements STEP et STAR de chaque établissement. Le code établissement sert à construire le NNT (Numéro National de Thèse)
- **date_soutenance :** date de soutenance de la thèse au format AAAA-MM-JJ. Pour les thèses provenant du SUDOC, seule l'année est significative
- **directeurs_these.[n].idref :** identifiant IdRef du directeur de la thèse
- **directeurs_these.[n].nom :** nom du directeur de la thèse
- **directeurs_these.[n].prenom :** prénom du directeur de la thèse
- **disciplines.fr :** discipline de la thèse
- **ecoles_doctorales.[n].idref :** identifiant IdRef de l'école doctorale
- **ecoles_doctorales.[n].nom :** nom de l'école doctorale
- **embargo :** date d'embargo de la thèse au format AAAA-MM-JJ
- **etablissements_soutenance.[n].idref :** identifiant IdRef de l'établissement de soutenance
- **etablissements_soutenance.[n].nom :** nom de l'établissement de soutenance
- **langue :** langue de la thèse au format ISO 639-1
- **membres_jury.[n].idref :** identifiant IdRef du membre du jury
- **membres_jury.[n].nom :** nom du membre du jury
- **membres_jury.[n].prenom :** prénom du membre du jury
- **nnt :** Numéro National de Thèse
- **oai_set_specs :** classification OAI de la thèse. **Au format CSV les codes sont séparés par ||.**
- **partenaires_recherche.[n].idref :** identifiant IdRef du partenaire de recherche
- **partenaires_recherche.[n].nom :** nom du partenaire de recherche
- **partenaires_recherche.[n].type :** type de partenaire de recherche (laboratoire, entreprise, équipe de recherche)
- **president_jury.idref :** identifiant IdRef du président du jury

Mise à jour mai 2022

- **president_jury.nom :** nom du président du jury
- **president_jury.prenom :** prénom du président du jury
- **rapporteurs.[n].idref :** identifiant IdRef du rapporteur
- **rapporteurs.[n].nom :** nom du rapporteur
- **rapporteurs.[n].prenom :** prénom du rapporteur
- **resumes.autre.[n] :** résumé de la thèse dans une langue autre que le français ou l'anglais. La langue du résumé est indiquée en début de chaîne de caractères par son code ISO 639-1 suivi de :: (ex. pour l'espagnol : es::). **Ce champ est spécifique au format CSV.**
- **resumes.en :** résumé de la thèse en anglais
- **resumes.fr :** résumé de la thèse en français
- **resumes.[x] :** résumé de la thèse dans la langue x. Par exemple pour l'espagnol resumes.es . **Ce champ est spécifique aux formats JSON et NDJSON.**
- **source :** origine des données (star, sudoc)
- **status :** statut de la thèse (soutenue, en cours)
- **sujets.autre.[n] :** sujet de la thèse dans une langue autre que le français ou l'anglais. La langue du titre est indiquée en début de chaîne de caractères par son code ISO 639-1 suivi de :: (ex. pour l'espagnol : es::). **Ce champ est spécifique au format CSV.**
- **sujets.en :** sujets de la thèse en anglais. **Au format CSV les sujets sont séparés par** ||
- **sujets.fr :** sujets de la thèse en français. **Au format CSV les sujets sont séparés par** ||
- **sujets.[x] :** sujet de la thèse dans la langue x. Par exemple pour l'espagnol sujets.es . **Ce champ est spécifique aux formats JSON et NDJSON.**
- **sujets_rameau :** indexation RAMEAU de la thèse. **Au format CSV les sujets sont séparés par ||.**
- **these_sur_travaux :** la thèse est une thèse sur travaux (oui, non)
- **titres.autre.[n] :** titre de la thèse dans une langue autre que l'anglais ou le français. La langue du titre est indiquée en début de chaîne de caractères par son code ISO 639-1 suivi de :: (ex. pour l'espagnol : es::). **Ce champ est spécifique au format CSV.**
- **titres.en :** titre de la thèse en anglais
- **titres.fr :** titre de la thèse en français
- **titres.[x]** : titre de la thèse dans la langue x. Par exemple pour l'espagnol titres.es . **Ce champ est spécifique aux formats JSON et NDJSON.**

Notes

Sur la notation [n]

La notation [n] indique que le champ est répétable.

Par exemple, si la thèse a deux auteurs, les informations concernant le premier auteur sont stockées dans :

- auteurs.0.idref
- auteurs.0.nom
- auteurs.0.prenom

et les informations concernant le second auteur sont stockées dans :

- auteurs.1.idref
- auteurs.1.nom
- auteurs.1.prenom

La numérotation [n] commence à 0.

Sur les identifiants IdRef et les tableurs

Les identifiants IdRef sont toujours composés de 9 caractères : 8 caractères numériques suivis d'un autre caractère numérique ou de X.

Les colonnes qui contiennent des identifiants IdRef doivent être impérativement au format texte. Ce formatage évite que les identifiants commençant par 0 soient raccourcis par le tableur.

Références

- Documentation de l'API de theses.fr
- Documentation de theses.fr
- Documentation de STEP
- Documentation de STAR
- Documentation d'IdRef
- Triple store data.idref.fr

Mise à jour mai 2022

abes

Guide réalisé par le Service des Thèses
Département Métadonnées et Services aux Réseaux - Contact : guichet d'assistance

Agence bibliographique de l'enseignement supérieur
avenue du Professeur Jean-Louis Viala - CS 84308 - 34193 Montpellier Cedex 5 - Tél. 04 67 54 84 10

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION