Istex pour les linguistes

Une ressource pour deux usages

Mathilde Huguin

Ingénieure INIST - CNRS Responsable du traitement et de l'analyse des données Istex
Jeune docteure associée au laboratoire ATILF

Journée des doctorants de l'Atilf, 8 décembre 2022

Plan

Introduction

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)
- Chargée d'enseignement 3 ans à l'Université de Lorraine

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)
- Chargée d'enseignement 3 ans à l'Université de Lorraine
- ATER 2 ans à l'Université de Lille & au laboratoire STL

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)
- Chargée d'enseignement 3 ans à l'Université de Lorraine
- ATER 2 ans à l'Université de Lille & au laboratoire STL
- Qualifiée aux fonctions de Maître de Conférences par le Conseil National des Universités pour la section 7

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)
- Chargée d'enseignement 3 ans à l'Université de Lorraine
- ATER 2 ans à l'Université de Lille & au laboratoire STL
- Qualifiée aux fonctions de Maître de Conférences par le Conseil National des Universités pour la section 7
- Ingénieure pour l'ANR Demonext *Dérivation en extension* (construction d'une base de données morphologiques)

Parcours

- Licence, Master et Doctorat à l'Université de Lorraine
- Thèse de morphologie : *Analyse morphologique des mots construits sur base de noms de personnalités politiques* sous la direction de Fiammetta Namer & Stéphanie Lignon (financée par un contrat du MESR)
- Chargée d'enseignement 3 ans à l'Université de Lorraine
- ATER 2 ans à l'Université de Lille & au laboratoire STL
- Qualifiée aux fonctions de Maître de Conférences par le Conseil National des Universités pour la section 7
- Ingénieure pour l'ANR Demonext *Dérivation en extension* (construction d'une base de données morphologiques)
- **Ingénieure à l'Inist : Responsable de l'analyse et du traitement des données Istex**

Objet de la présentation

1. Découvrir **l'Inist**, le travail d'un ingénieur

Objet de la présentation

1. Découvrir **l'Inist**, le travail d'un ingénieur
2. Découvrir **Istex** et les outils associés

Qu’est-ce que l’Inist ?

- *Institut de l’Information Scientifique et Technique*

Qu’est-ce que l’Inist ?

- *Institut de l’Information Scientifique et Technique*
- Unité d’Appui et de Recherche du CNRS (UAR76)

Qu’est-ce que l’Inist ?

- *Institut de l’Information Scientifique et Technique*
- Unité d’Appui et de Recherche du CNRS (UAR76)
- Institut : Direction Générale Déléguée à la Science (DGDS)

Qu'est-ce que l'Inist ?

- *Institut de l'Information Scientifique et Technique*
- Unité d'Appui et de Recherche du CNRS (UAR76)
- Institut : Direction Générale Déléguée à la Science (DGDS)
- Direction : Direction des Données Ouvertes de la Recherche (DDOR)

Qu’est-ce que l’Inist ?

- **Mission** : faciliter l’accès, l’analyse et la fouille de l’information scientifique et valoriser la production scientifique

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets

Qu’est-ce que l’Inist ?

- **Mission** : faciliter l’accès, l’analyse et la fouille de l’information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)

Qu’est-ce que l’Inist ?

- **Mission** : faciliter l’accès, l’analyse et la fouille de l’information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d’**Istex** (archives scientifiques)
  - Extraction de données et création d’**outils de TDM**

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)
  - Extraction de données et création d'**outils de TDM**
    - Thésaurus de paléoclimatologie : 2000 concepts pour offrir un vocabulaire de référence et partager des données de recherche

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)
  - Extraction de données et création d'**outils de TDM**
    - Thésaurus de paléoclimatologie : 2000 concepts pour offrir un vocabulaire de référence et partager des données de recherche
    - TDM : outil de visualisation **Lodex** (Linked Open Data EXperiment)

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)
  - Extraction de données et création d'**outils de TDM**
    - Thésaurus de paléoclimatologie : 2000 concepts pour offrir un vocabulaire de référence et partager des données de recherche
    - TDM : outil de visualisation **Lodex** (Linked Open Data EXperiment)
  - **Formations** sur la gestion et le partage de données (aspects juridiques, dépôts, identifiants pérennes, fouille de texte)

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)
  - Extraction de données et création d'**outils de TDM**
    - Thésaurus de paléoclimatologie : 2000 concepts pour offrir un vocabulaire de référence et partager des données de recherche
    - TDM : outil de visualisation **Lodex** (Linked Open Data EXperiment)
  - **Formations** sur la gestion et le partage de données (aspects juridiques, dépôts, identifiants pérennes, fouille de texte)
  - **Bibliométrie** (pour les rapports HCERES, les bibliothèques universitaires)

Qu'est-ce que l'Inist ?

- **Mission** : faciliter l'accès, l'analyse et la fouille de l'information scientifique et valoriser la production scientifique
- Quelques exemples concrets
  - Création de **BibCnrs** ou modération de **HAL** (publications courantes)
  - Création et enrichissement d'**Istex** (archives scientifiques)
  - Extraction de données et création d'**outils de TDM**
    - Thésaurus de paléoclimatologie : 2000 concepts pour offrir un vocabulaire de référence et partager des données de recherche
    - TDM : outil de visualisation **Lodex** (Linked Open Data EXperiment)
  - **Formations** sur la gestion et le partage de données (aspects juridiques, dépôts, identifiants pérennes, fouille de texte)
  - **Bibliométrie** (pour les rapports HCERES, les bibliothèques universitaires)
  - Service de **traduction** (pensez-y !)

Présentation d’Istex

Qu’est-ce qu’Istex ?

- *Initiative d’excellence en Information Scientifique et Technique*

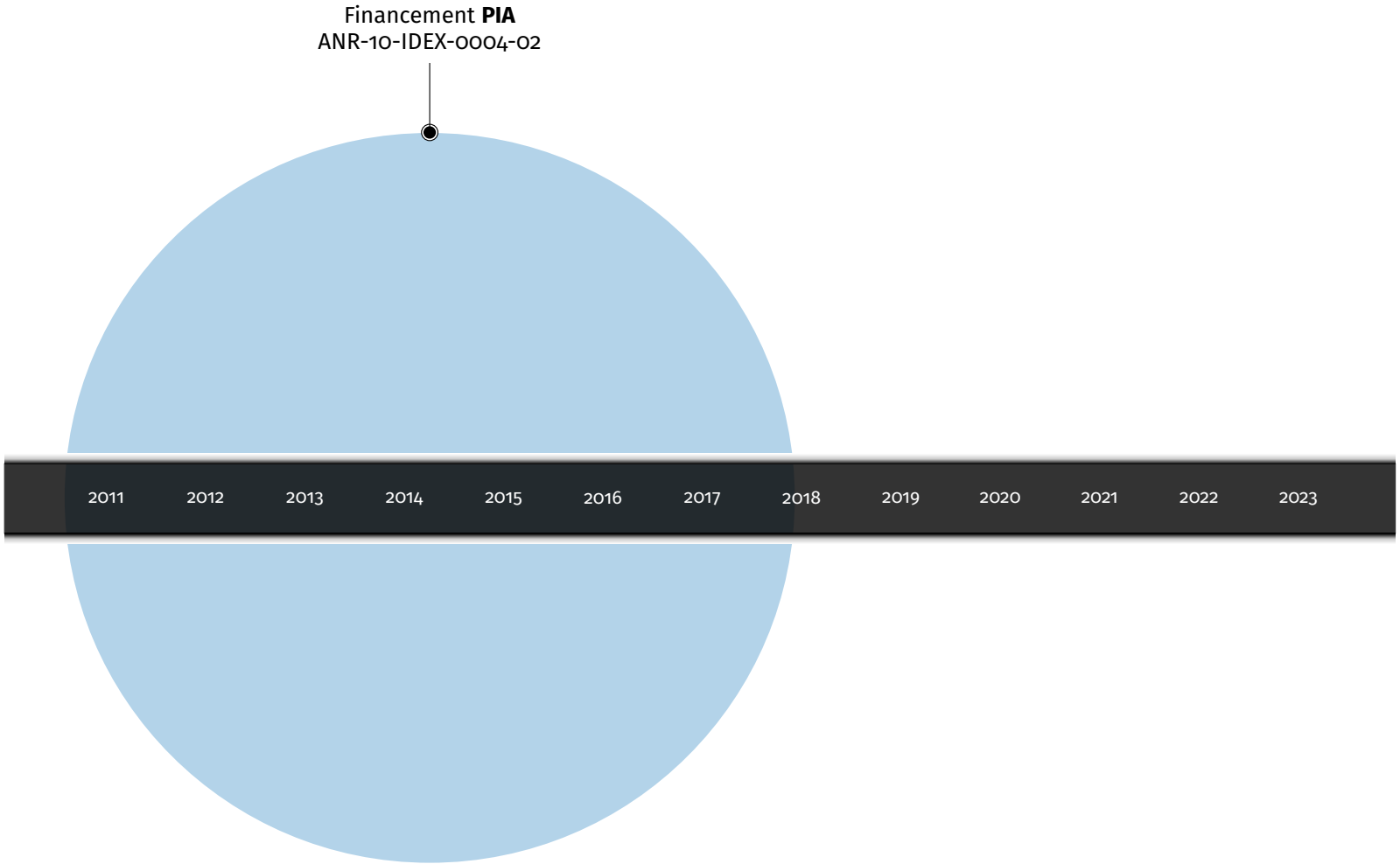
Qu'est-ce qu'Istex ?

- *Initiative d'excellence en Information Scientifique et Technique*
- Istex (2011-2018) est à l'origine un projet qui, dans le cadre des PIA, se donne comme objectif de **construire une bibliothèque numérique d'archives scientifiques**

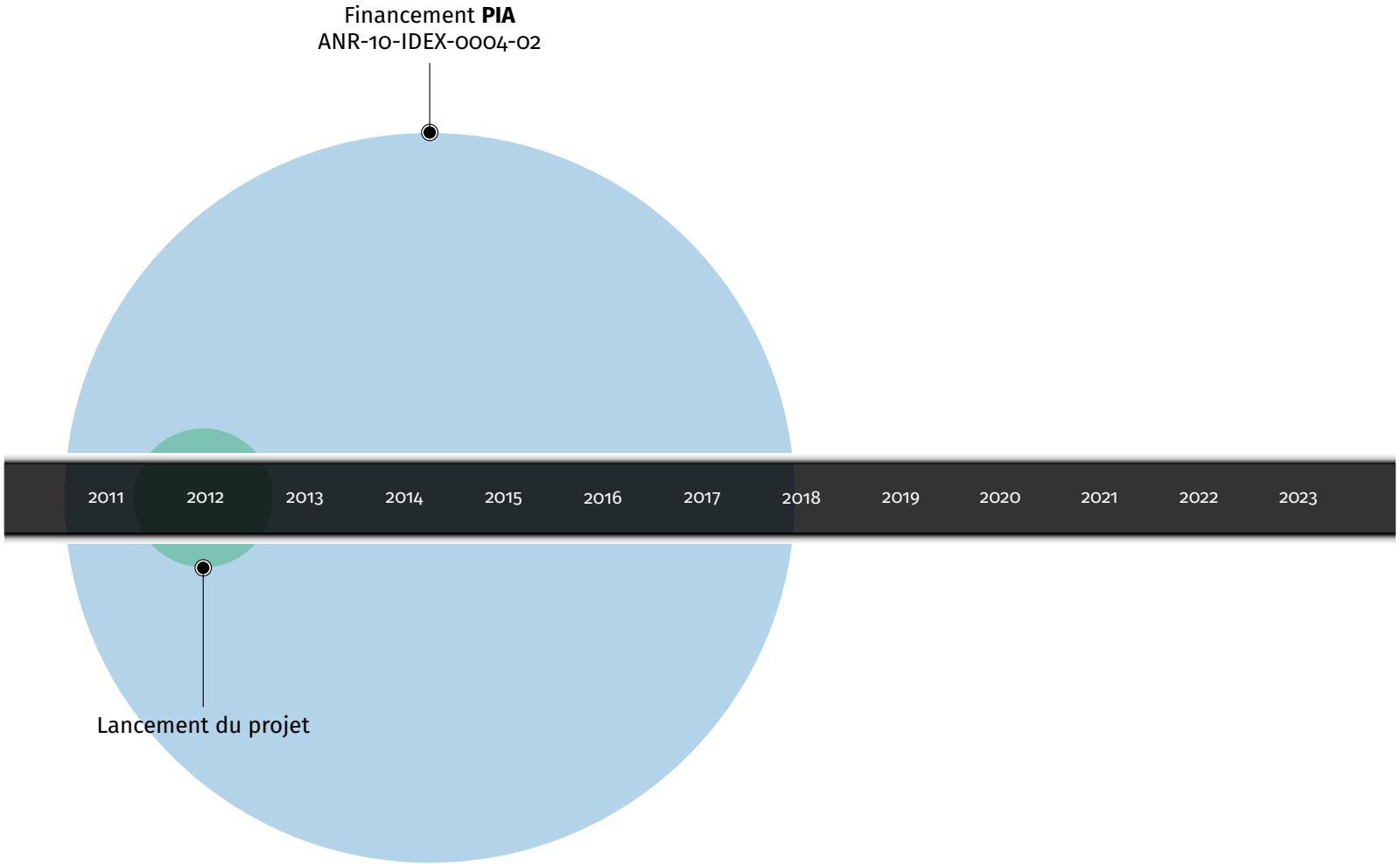
Qu'est-ce qu'Istex ?

- *Initiative d'excellence en Information Scientifique et Technique*
- Istex (2011-2018) est à l'origine un projet qui, dans le cadre des PIA, se donne comme objectif de **construire une bibliothèque numérique d'archives scientifiques**
- Après l'obtention de licences nationales auprès des éditeurs scientifiques, les publications sont mises à disposition des personnels et étudiants de l'ESR *via* la plateforme Istex

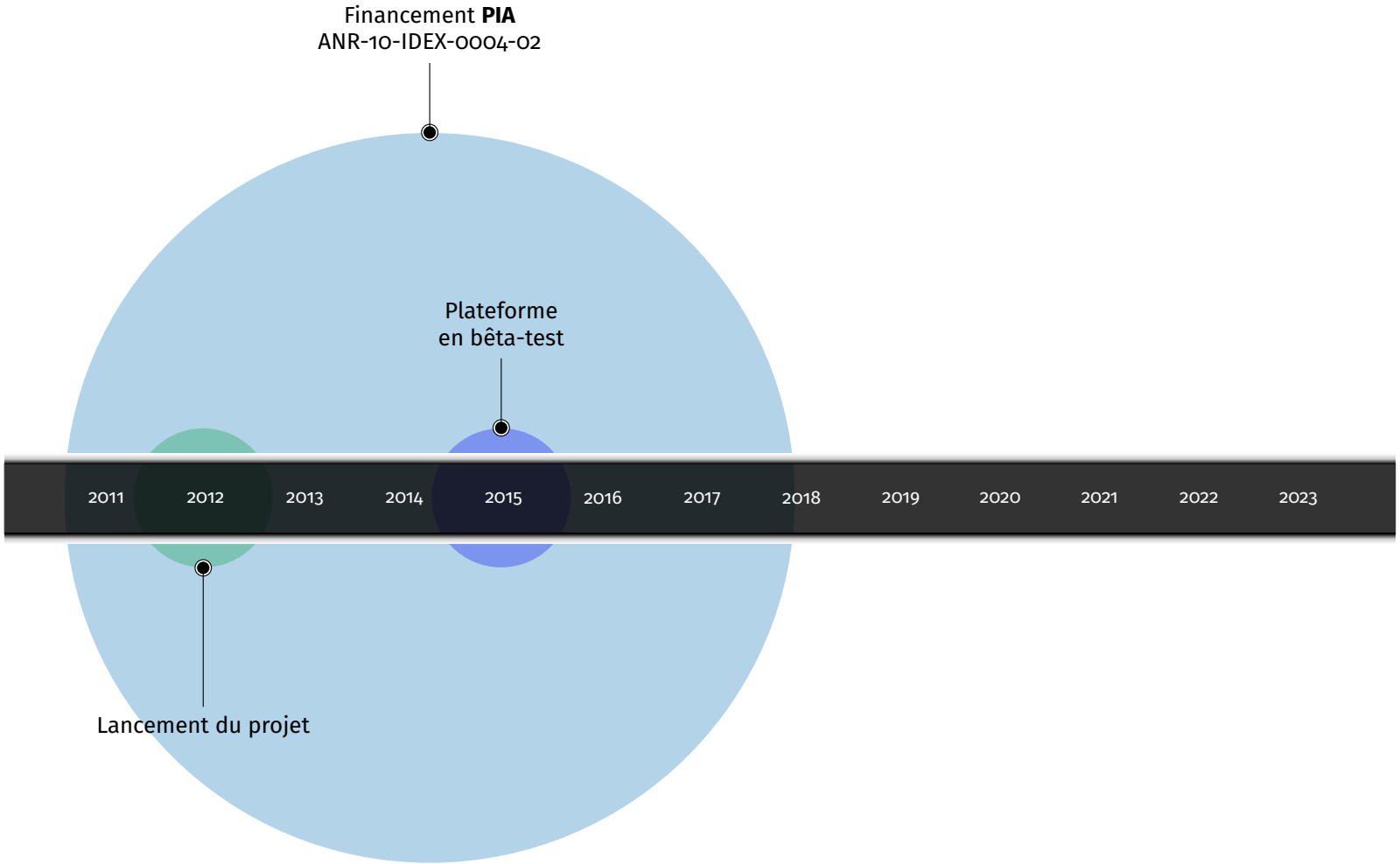
Qu’est-ce qu’Istex ?



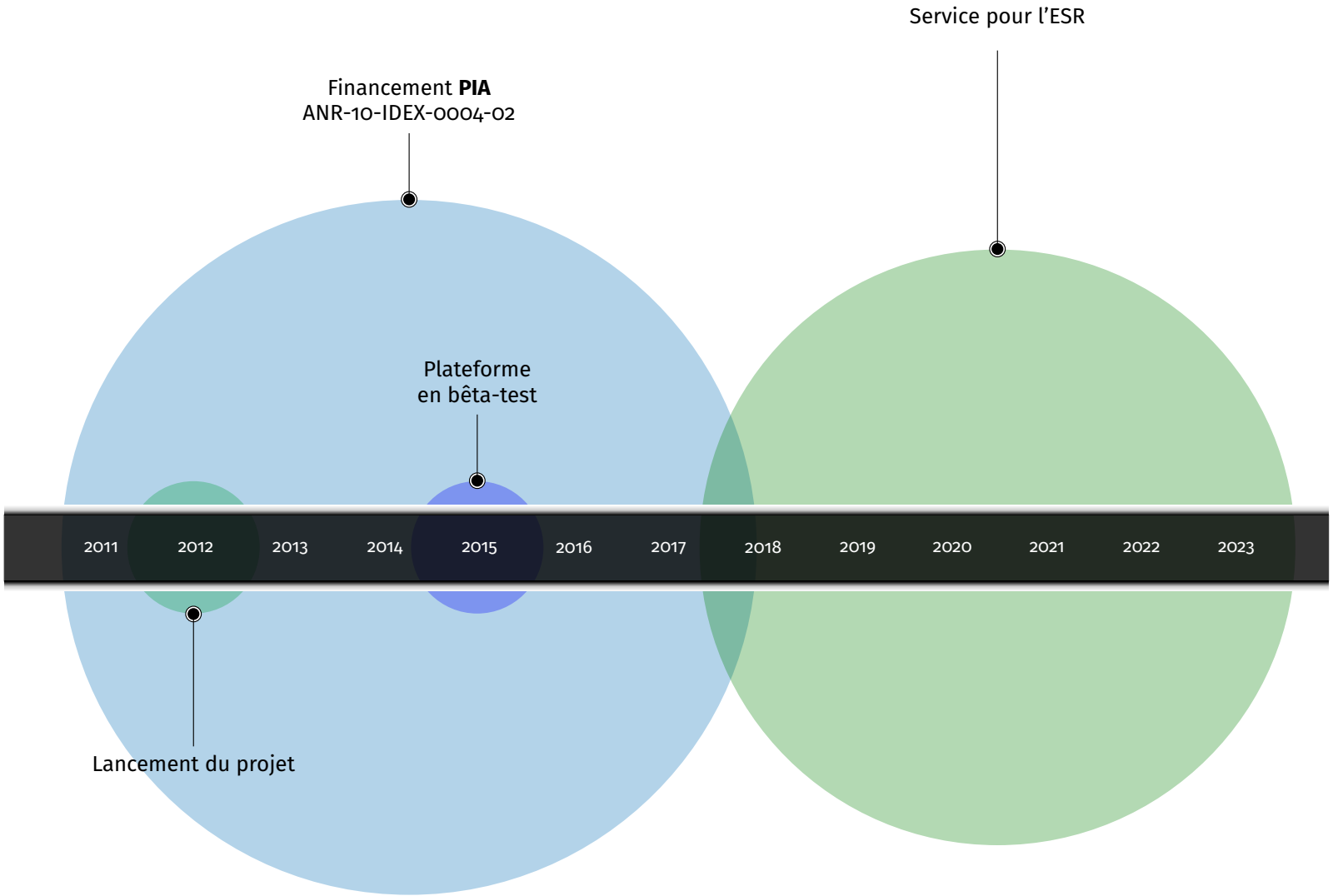
Qu’est-ce qu’Istex ?



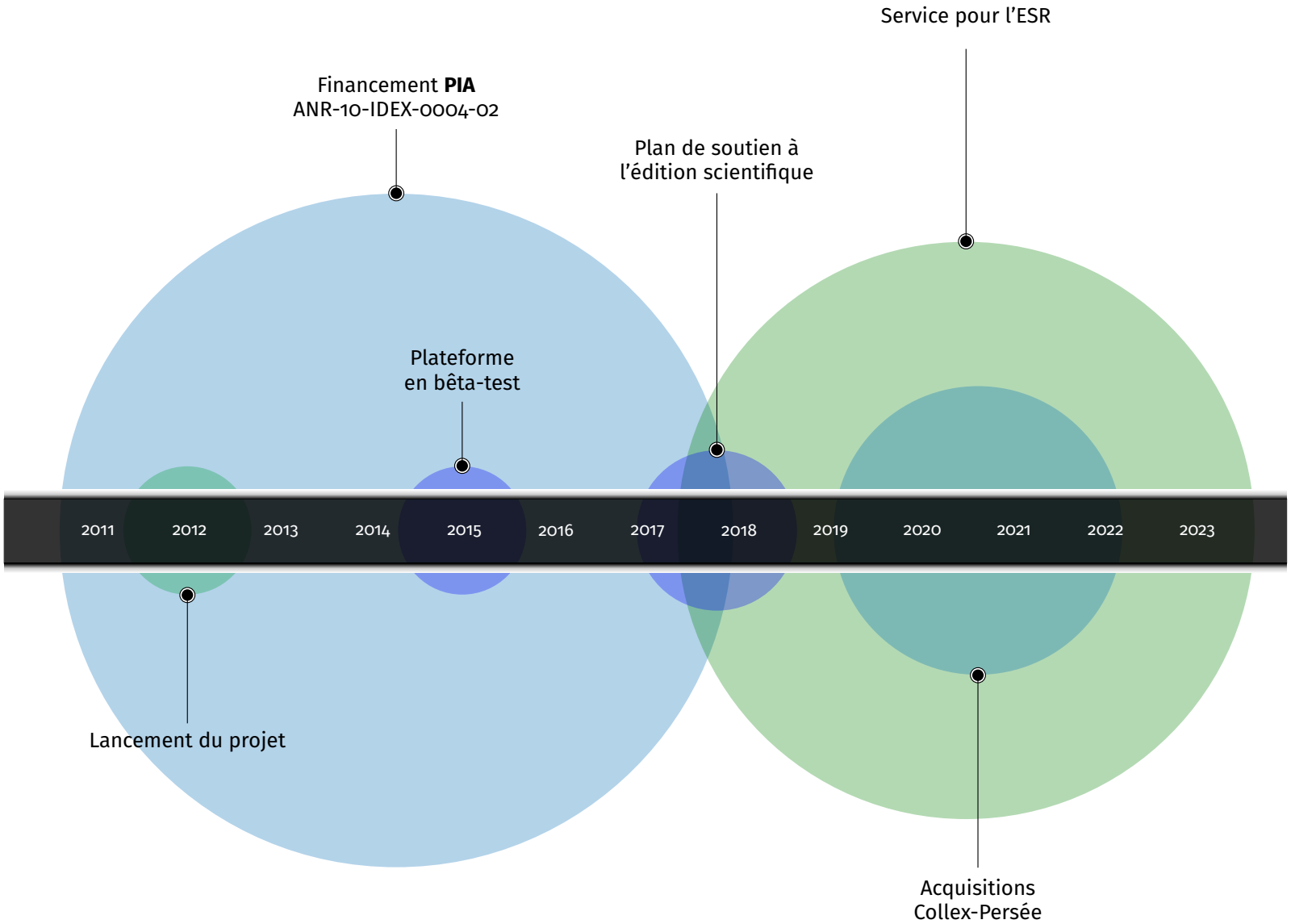
Qu’est-ce qu’Istex ?



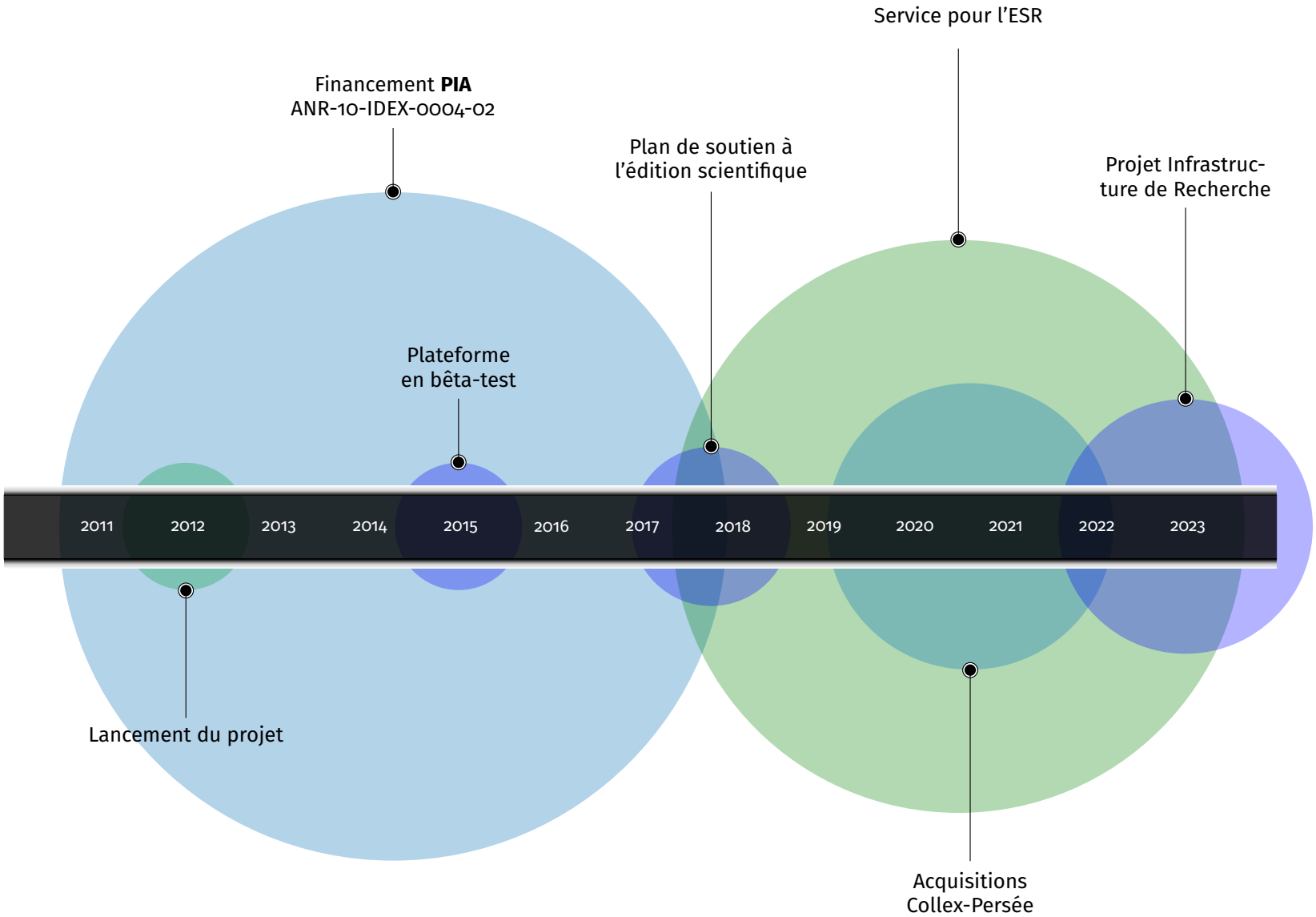
Qu’est-ce qu’Istex ?



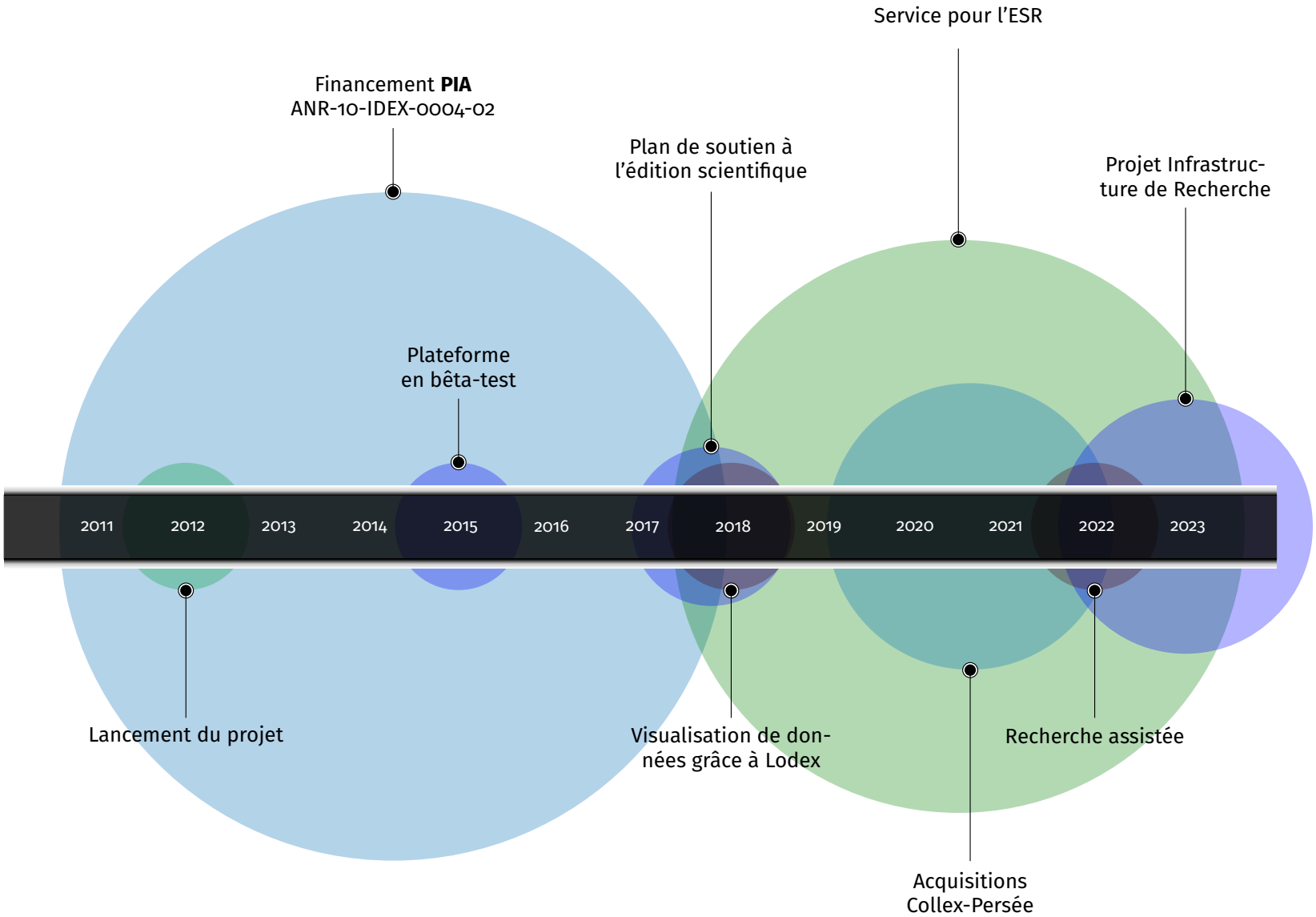
Qu’est-ce qu’Istex ?



Qu’est-ce qu’Istex ?



Qu’est-ce qu’Istex ?



Qu’est-ce qu’Istex ?

Un "Frantext scientifique"

- Un réservoir d’articles scientifiques, de reviews ou d’ebooks **multilingues** et **multidisciplinaires**

Qu’est-ce qu’Istex ?

Un "Frantext scientifique"

- Un réservoir d’articles scientifiques, de reviews ou d’ebooks **multilingues** et **multidisciplinaires**
- En constante évolution, dernier ajout : 3 500 documents « Théologiens chrétiens du XXe siècle »

Qu’est-ce qu’Istex ?

Un "Frantext scientifique"

- Un réservoir d’articles scientifiques, de reviews ou d’ebooks **multilingues** et **multidisciplinaires**
- En constante évolution, dernier ajout : 3 500 documents « Théologiens chrétiens du XXe siècle »
- Accessible à tous les personnels de l’ESR (lien avec des ENT, des référentiels comme BibCnrs)

Quelques chiffres*

*Chiffres en date du 8 décembre 2022

Quelques chiffres*

27 282 527 documents

*Chiffres en date du 8 décembre 2022

Quelques chiffres*

27 282 527 documents

38 collections d’éditeurs

*Chiffres en date du 8 décembre 2022

Quelques chiffres*

27 282 527 documents

38 collections d’éditeurs

9 318 revues

*Chiffres en date du 8 décembre 2022

Quelques chiffres*

27 282 527 documents

38 collections d’éditeurs

9 318 revues

353 710 monographies

*Chiffres en date du 8 décembre 2022

Quelques chiffres*

27 282 527 documents

38 collections d'éditeurs

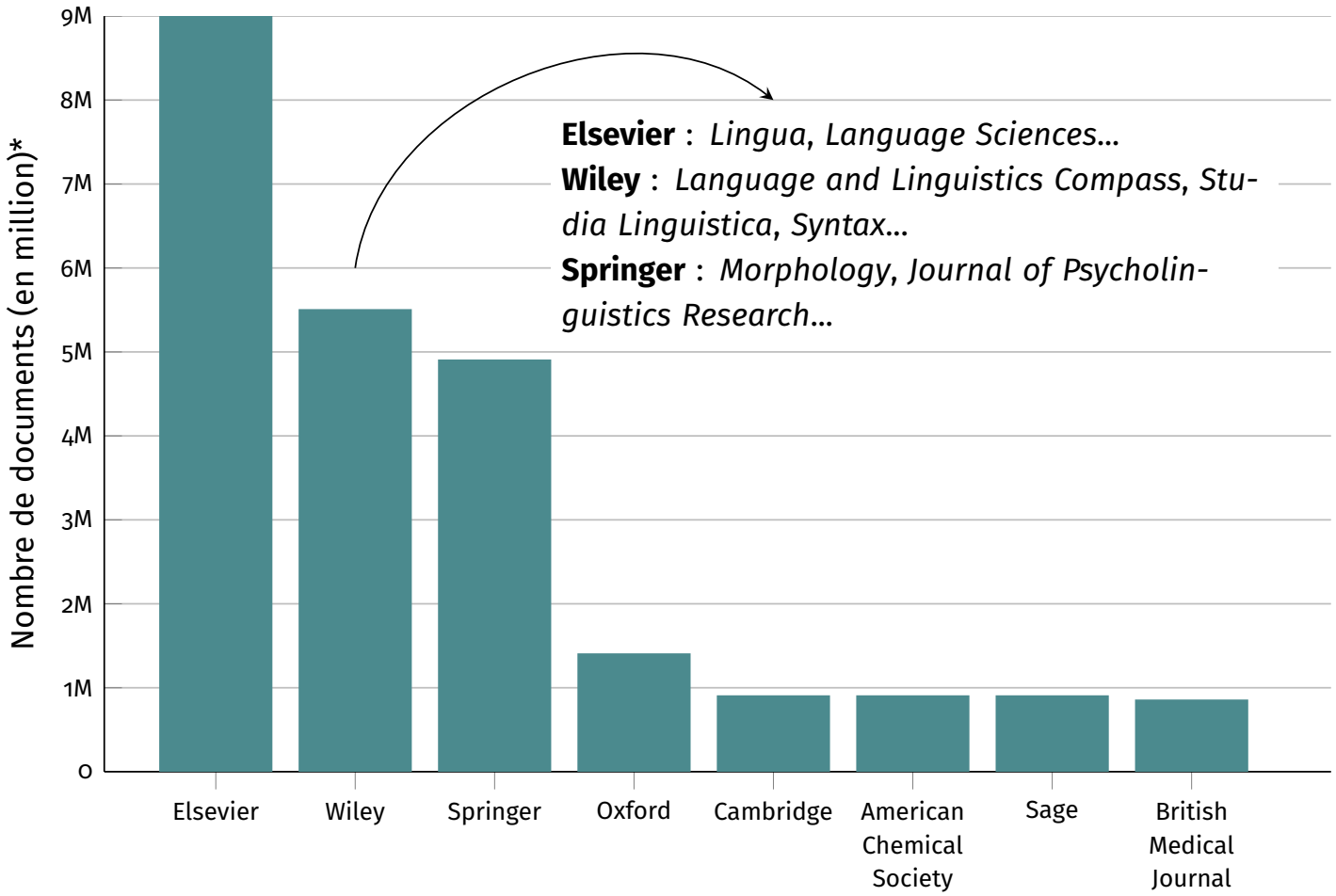
9 318 revues

353 710 monographies

700 ans de publications

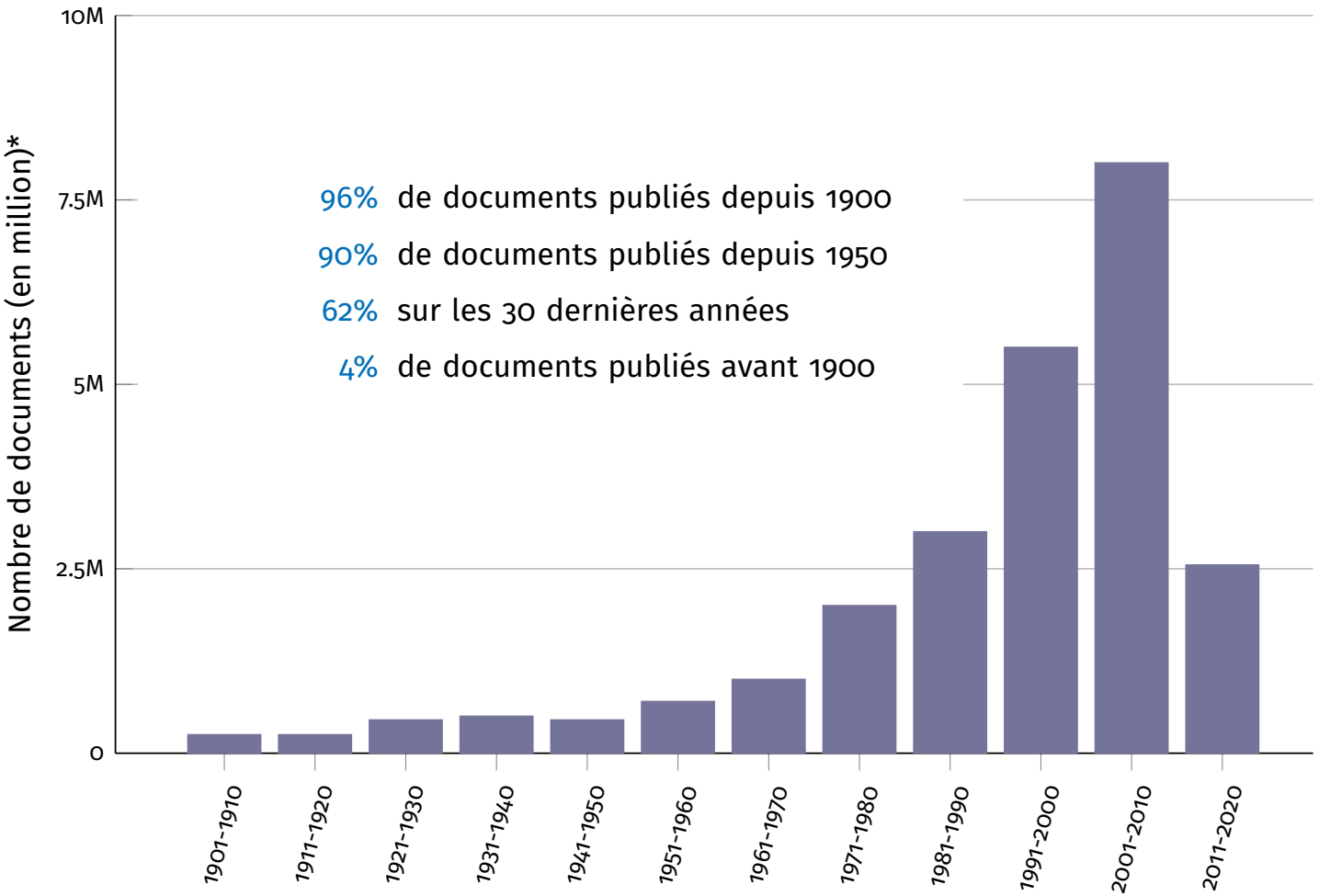
*Chiffres en date du 8 décembre 2022

Principaux éditeurs



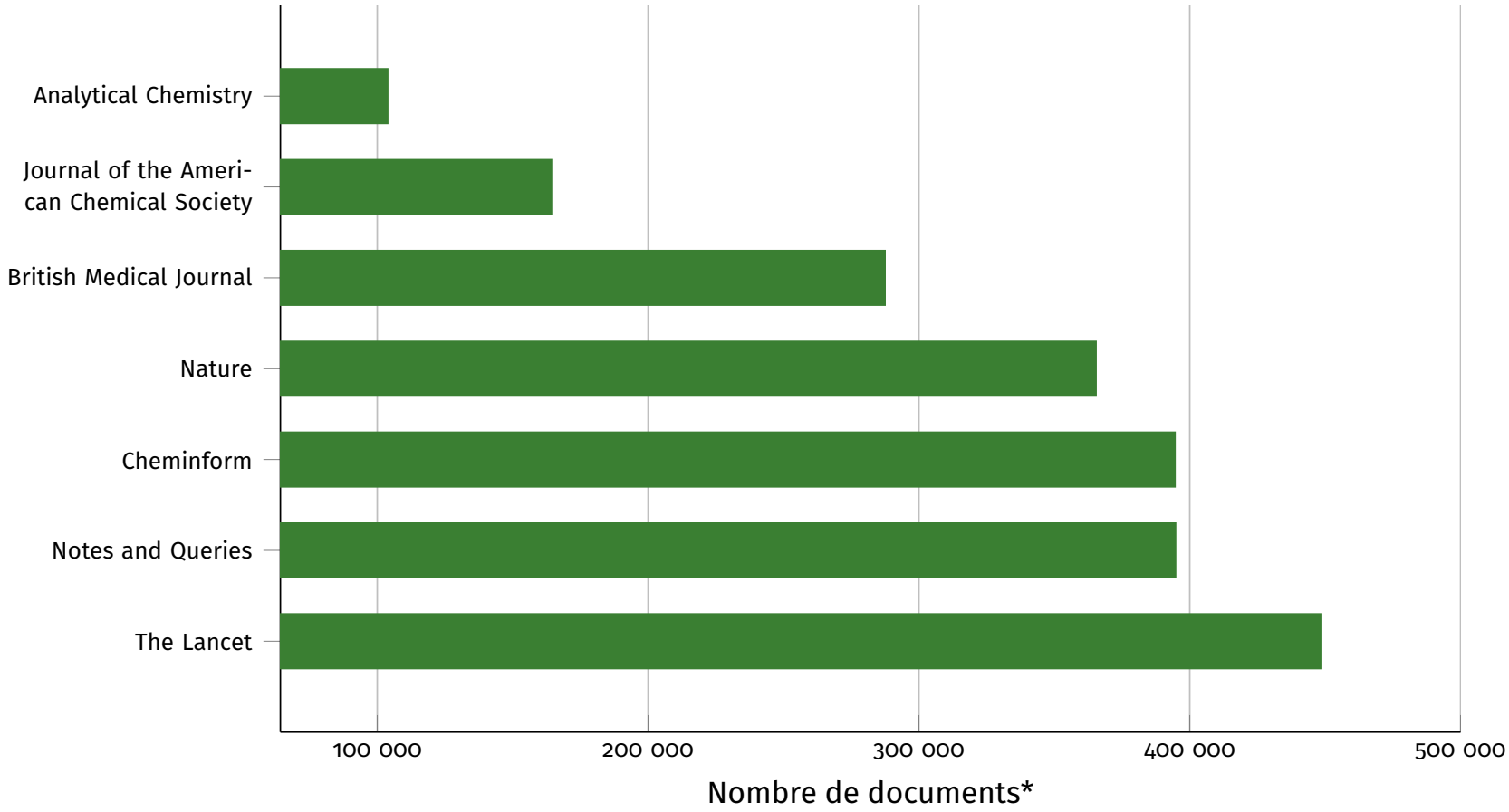
*Chiffres en date du 8 décembre 2022

700 ans de publication



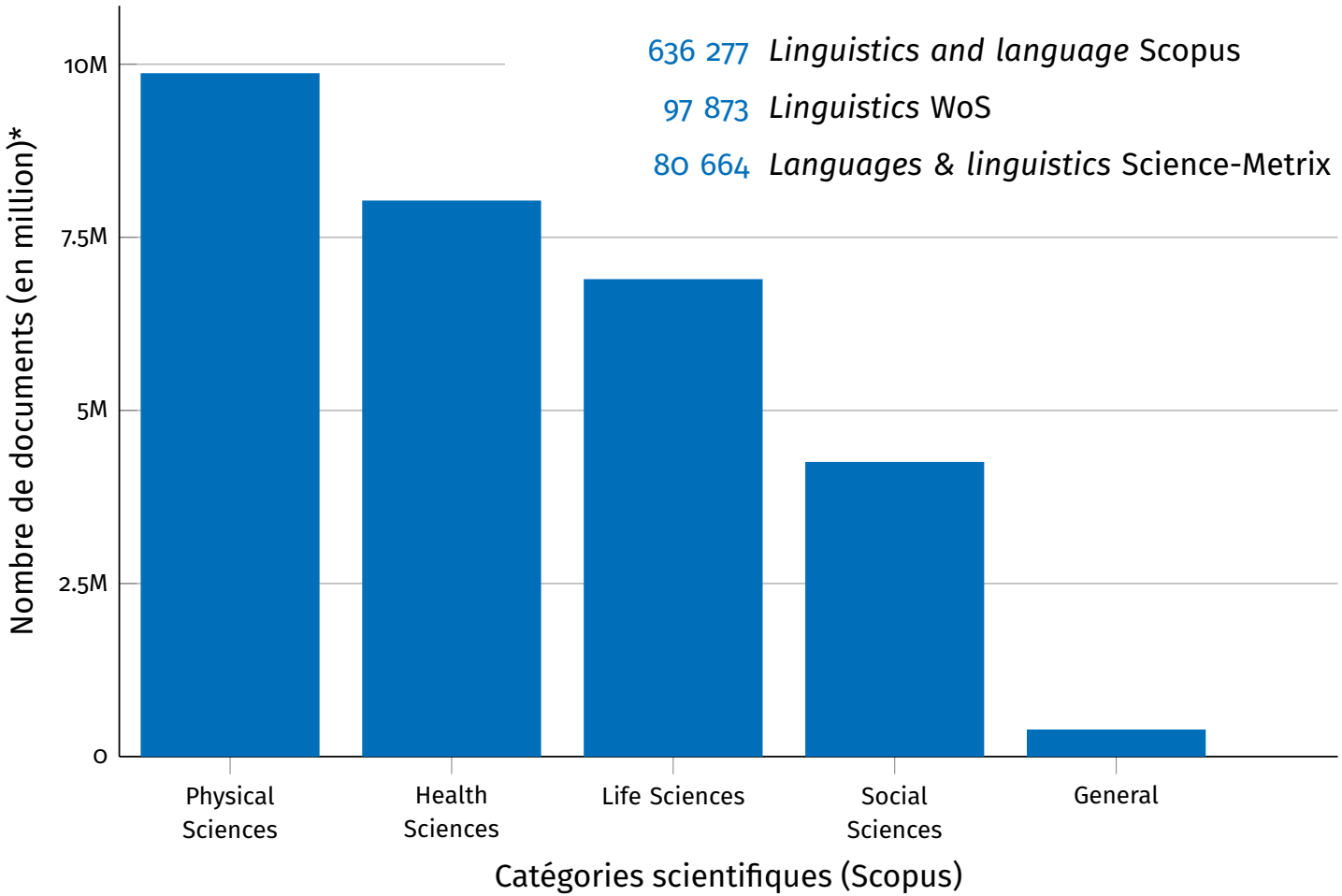
*Chiffres en date du 8 décembre 2022

Principales revues



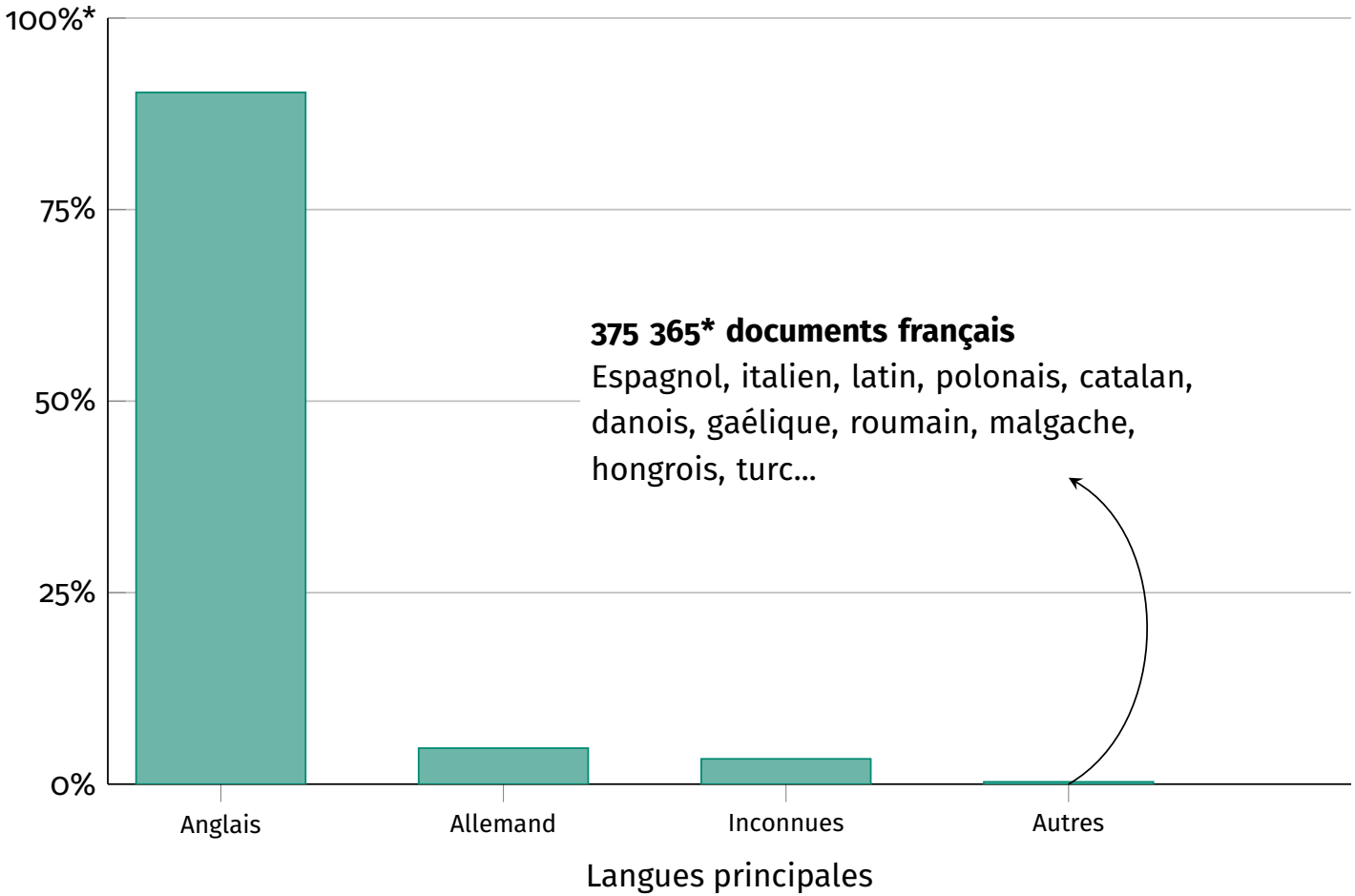
*Chiffres en date du 8 décembre 2022

Tous les domaines scientifiques



*Chiffres en date du 8 décembre 2022

Une ressource multilingue : 51 langues



*Chiffres en date du 8 décembre 2022

Pour quoi faire ?

1. Une **ressource documentaire** pour trouver des articles non accessibles chez l'éditeur

Pour quoi faire ?

1. Une **ressource documentaire** pour trouver des articles non accessibles chez l'éditeur
2. Un **matériau de recherche** pour constituer des corpus d'articles scientifiques et donc faire de la linguistique de corpus, de la fouille de texte, du TAL

Pour quoi faire ?

1. Une **ressource documentaire** pour trouver des articles non accessibles chez l'éditeur
2. Un **matériau de recherche** pour constituer des corpus d'articles scientifiques et donc faire de la linguistique de corpus, de la fouille de texte, du TAL
    - Étude du discours scientifique, étude en terminologie, étude diachronique, linguistique comparée, étude de langues rares, ressources pour exemples authentiques en syntaxe, ressource pour enrichir des bases de données lexicales avec des contextes authentiques...

Pour quoi faire ?

1. Une **ressource documentaire** pour trouver des articles non accessibles chez l'éditeur
2. Un **matériau de recherche** pour constituer des corpus d'articles scientifiques et donc faire de la linguistique de corpus, de la fouille de texte, du TAL
   - Étude du discours scientifique, étude en terminologie, étude diachronique, linguistique comparée, étude de langues rares, ressources pour exemples authentiques en syntaxe, ressource pour enrichir des bases de données lexicales avec des contextes authentiques...
   - Dernier exemple en date : *Researchers and their data. A study based on the use of the word data in scholarly articles* (Maisonobe et Bordignon, 2022)

Les outils

- Istex.fr : le site général vous renvoie vers différents outils

Les outils

- Istex.fr : le site général vous renvoie vers différents outils
- data.istex.fr : permet d'accéder à toutes les **informations** concernant le contenu d'Istex, ses collections, ses enrichissements, les corpus déjà construits

Les outils

- Istex.fr : le site général vous renvoie vers différents outils
- data.istex.fr : permet d'accéder à toutes les **informations** concernant le contenu d'Istex, ses collections, ses enrichissements, les corpus déjà construits
- Istex-DL : permet d'**écrire une requête** et de **télécharger** un corpus (on peut aussi le faire directement *via* l'API)

Les outils

- Istex.fr : le site général vous renvoie vers différents outils
- data.istex.fr : permet d'accéder à toutes les **informations** concernant le contenu d'Istex, ses collections, ses enrichissements, les corpus déjà construits
- Istex-DL : permet d'**écrire une requête** et de **télécharger** un corpus (on peut aussi le faire directement *via* l'API)
- Lodex : un logiciel open source dédié **à la valorisation de données structurées** (exemple des étudiants de l'Enssib)

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais
   - Structuration des pdf (Grobid) : titre, auteur, affiliation, mots-clés, résumé, corps du texte, références bibliographiques

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais
   - Structuration des pdf (Grobid) : titre, auteur, affiliation, mots-clés, résumé, corps du texte, références bibliographiques
   - Indexation (Teeft) : mots-clés des articles en anglais

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais
   - Structuration des pdf (Grobid) : titre, auteur, affiliation, mots-clés, résumé, corps du texte, références bibliographiques
   - Indexation (Teeft) : mots-clés des articles en anglais
   - Catégorisations par domaines scientifiques (Multicat) : différents référentiels (Inist, Science-Metrix, Scopus, WoS)

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais
   - Structuration des pdf (Grobid) : titre, auteur, affiliation, mots-clés, résumé, corps du texte, références bibliographiques
   - Indexation (Teeft) : mots-clés des articles en anglais
   - Catégorisations par domaines scientifiques (Multicat) : différents référentiels (Inist, Science-Metrix, Scopus, WoS)
   - Caractérisation des documents : nombre de mots, score de qualité, présence et type d'enrichissements

Données et enrichissements

Focus sur la chaîne de traitement : la valeur ajoutée d'Istex

1. Réception
2. Stockage, standardisation et indexation (océrisation des PDF si absence d'information textuelle, TEI *Text Encoding Initiative*)
3. Enrichissements
   - Entités nommées (Unitex) : anthroponymes, toponymes, ergonymes... en français et en anglais
   - Structuration des pdf (Grobid) : titre, auteur, affiliation, mots-clés, résumé, corps du texte, références bibliographiques
   - Indexation (Teeft) : mots-clés des articles en anglais
   - Catégorisations par domaines scientifiques (Multicat) : différents référentiels (Inist, Science-Metrix, Scopus, WoS)
   - Caractérisation des documents : nombre de mots, score de qualité, présence et type d'enrichissements
4. Textes intégraux à valeur ajoutée téléchargeables massivement

Construire un corpus

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet
- **Méthode**

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet
- **Méthode**
  1. Élaboration d'une requête pour sélectionner les documents

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet
- **Méthode**
  1. Élaboration d'une requête pour sélectionner les documents
  2. Téléchargement des documents

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet
- **Méthode**
  1. Élaboration d'une requête pour sélectionner les documents
  2. Téléchargement des documents
  3. Exploration des données

Présentation de l'exemple

- Travail sur un corpus correspondant à un sujet de thèse : *Analyse morphologique des mots construits sur base de noms de personnalités politiques*
- **Objectif du corpus** : permettre l'identification de thèmes récurrents, de notions clés, d'auteurs indispensables ou encore de critères définitoires nécessaires pour débuter la recherche sur ce sujet
- **Méthode**
  1. Élaboration d'une requête pour sélectionner les documents
  2. Téléchargement des documents
  3. Exploration des données
- Cette méthode est **itérative** : l'exploration des résultats fait détecter le bruit qui amène éventuellement à réviser la requête initiale

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)
  - En utilisant **Istex-DL** et sa recherche assistée qui permettent d'accéder à l'API dans une interface plus ergonomique

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)
  - En utilisant **Istex-DL** et sa recherche assistée qui permettent d'accéder à l'API dans une interface plus ergonomique
- Les termes de la requête du corpus de thèse

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)
  - En utilisant **Istex-DL** et sa recherche assistée qui permettent d'accéder à l'API dans une interface plus ergonomique
- Les termes de la requête du corpus de thèse
  - Nom propre : différentes graphies possibles de *nom propre* en français et en anglais

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)
  - En utilisant **Istex-DL** et sa recherche assistée qui permettent d'accéder à l'API dans une interface plus ergonomique
- Les termes de la requête du corpus de thèse
  - Nom propre : différentes graphies possibles de *nom propre* en français et en anglais
  - Morphologie : différentes dénominations possibles des opérations (*suffixation*), des unités (*lexème*) ou encore des sous-domaines de la morphologie (*morpho-sémantique*), en français et en anglais

Élaborer une requête

- Le requêtage d'Istex utilise une syntaxe particulière (Lucene, opérateurs booléens, Regex)
- Deux moyens d'accéder aux documents
  - En interrogeant directement l'**API** : `https://api.istex.fr/document/?q=` (« interface de programmation applicative », i.e. un système ou un ensemble de composants permettant d'interagir avec un service)
  - En utilisant **Istex-DL** et sa recherche assistée qui permettent d'accéder à l'API dans une interface plus ergonomique
- Les termes de la requête du corpus de thèse
  - Nom propre : différentes graphies possibles de *nom propre* en français et en anglais
  - Morphologie : différentes dénominations possibles des opérations (*suffixation*), des unités (*lexème*) ou encore des sous-domaines de la morphologie (*morpho-sémantique*), en français et en anglais
- Les champs choisis : titre, résumé et mots-clés d'auteur(s)

Élaborer une requête

Démonstrations dans l'API et dans Istex-DL

- `https://api.istex.fr/document/?q=`
- `https://dl.istex.fr/`

Élaborer une requête

(title.raw:(/.*[pP]roper.[nN]ames?.*/ /.*[pP]roper.[nN]ouns?.*/ /.*[Nn]oms ?.[pP]ropres?.*/) OR abstract.raw:(/.*[pP]roper.[nN]ames?.*/ /.*[pP]roper.[nN]ouns?.*/ /.*[Nn]oms ?.[pP]ropres?.*/) OR subject.value.raw:(/.*[pP]roper.[nN]ames?.*/ /.*[pP]roper.[nN]ouns?.*/ /.*[Nn]oms ?.[pP]ropres?.*/)) AND publicationDate:[1923 TO 2019] NOT categories.wos.raw:("2 - neurosciences" "2 - gerontology" "2 - computer science, interdisciplinary applications") NOT categories.inist.raw:("3 - sciences medicales") NOT categories.scopus.raw:("2 - Computer Science" "3 - Archaeology" "2 - Physics and Astronomy" "2 - Environmental Science" "2 - Earth and Planetary Sciences" "2 - Immunology and Microbiology" "2 - Agricultural and Biological Sciences" "2 - Nursing" "2 - Veterinary" "3 - Radiology Nuclear Medicine and imaging" "3 - Infectious Diseases" "3 - Public Health, Environmental and Occupational Health" "3 - Medicine (miscellaneous)") NOT categories.scienceMetrix.raw :("2 - agriculture, fisheries & forestry" "2 - visual & performing arts" "2 - engineering" "2 - economics & business" "3 - criminology" "3 - cultural studies" "3 - geography" "3 - information & library sciences" "3 - international relations" "3 - law" "3 - science studies" "3 - social work" "3 - mycology & parasitology" "3 - anatomy & morphology" "3 - endocrinology & metabolism" "3 - general & internal medicine" "3 - neurology& neurosurgery" "3 - tropical medicine" "2 - public health & health services" "2 - physics & astronomy" "3 - statistics & probability" "2 - biology" "2 - earth & environmental sciences") NOT /onomasti[cq].*/ NOT corpusName.raw:("eebo" "ecco") NOT host.title.raw :("Notes and Queries") NOT title:("index" "errata")

Comment télécharger un corpus ?

- Istex-DL permet de télécharger 100 000 documents en choisissant le **format** qui dépend des post-traitements envisagés (Lodex, TXM, Excel, etc.)

Comment télécharger un corpus ?

- Istex-DL permet de télécharger 100 000 documents en choisissant le **format** qui dépend des post-traitements envisagés (Lodex, TXM, Excel, etc.)
- On peut aussi choisir les **données** que l'on télécharge : le texte intégral, les références bibliographiques, les domaines scientifiques, les mots-clés…

Explorer et visualiser

- Lodex est un outil open-source associant **sémantisation et visualisation de données**

Explorer et visualiser

- Lodex est un outil open-source associant **sémantisation et visualisation de données**
- L'outil permet de créer des sites web offrant des interfaces pour explorer et naviguer dans un jeu de données au travers d'une liste de fiches ou d'une série de graphiques dynamiques (histogrammes, cartes, diachronies, etc.)

Explorer et visualiser

- Lodex est un outil open-source associant **sémantisation et visualisation de données**
- L'outil permet de créer des sites web offrant des interfaces pour explorer et naviguer dans un jeu de données au travers d'une liste de fiches ou d'une série de graphiques dynamiques (histogrammes, cartes, diachronies, etc.)
- Vous pouvez l'utiliser pour **n'importe quel autre jeu de données structurées**

Explorer et visualiser

- Lodex est un outil open-source associant **sémantisation et visualisation de données**
- L'outil permet de créer des sites web offrant des interfaces pour explorer et naviguer dans un jeu de données au travers d'une liste de fiches ou d'une série de graphiques dynamiques (histogrammes, cartes, diachronies, etc.)
- Vous pouvez l'utiliser pour **n'importe quel autre jeu de données structurées**
- Vous avez accès à un back-office qui contient vos données et vous sélectionnez différents affichages, graphiques et méthodes de filtres qui seront visibles sur la page publiée

Les résultats

Exploration du corpus grâce à Lodex

- Chargement des données téléchargées et aperçu du travail sur Lodex
- Résultats du corpus d'exemple *Nom propre et Morphologie* : `https://atilf-phd-1.lodex.inist.fr/`

Conclusion

- Utilisez Istex !

Conclusion

- Utilisez Istex !
- N'hésitez pas à contacter l'Inist si vous avez besoin d'outils ou d'aide pour utiliser les services proposés !

Conclusion

- Utilisez Istex !
- N'hésitez pas à contacter l'Inist si vous avez besoin d'outils ou d'aide pour utiliser les services proposés !
  - *Via* le formulaire : `https://www.istex.fr/contact/`

Conclusion

- Utilisez Istex !
- N'hésitez pas à contacter l'Inist si vous avez besoin d'outils ou d'aide pour utiliser les services proposés !
  - *Via* le formulaire : `https://www.istex.fr/contact/`
  - *Via* la liste : `contact@listes.istex.fr`

Conclusion

- Utilisez Istex !
- N'hésitez pas à contacter l'Inist si vous avez besoin d'outils ou d'aide pour utiliser les services proposés !
  - *Via* le formulaire : `https://www.istex.fr/contact/`
  - *Via* la liste : `contact@listes.istex.fr`
  - `mathilde.huguin@inist.fr`

Conclusion

- Utilisez Istex !
- N'hésitez pas à contacter l'Inist si vous avez besoin d'outils ou d'aide pour utiliser les services proposés !
  - *Via* le formulaire : `https://www.istex.fr/contact/`
  - *Via* la liste : `contact@listes.istex.fr`
  - `mathilde.huguin@inist.fr`
- A venir très prochainement : des articles de SHS de CAIRN en français avec du texte en très bonne qualité !

Merci de votre attention !