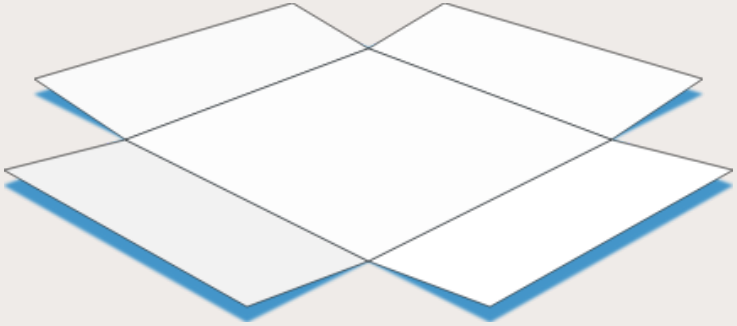


Produire, ouvrir et valoriser les logiciels de recherche

Retour d'expérience en SHS : Hyphe

Colloque Sciences Ouvertes 2024 – Université de Lorraine

28 novembre 2024

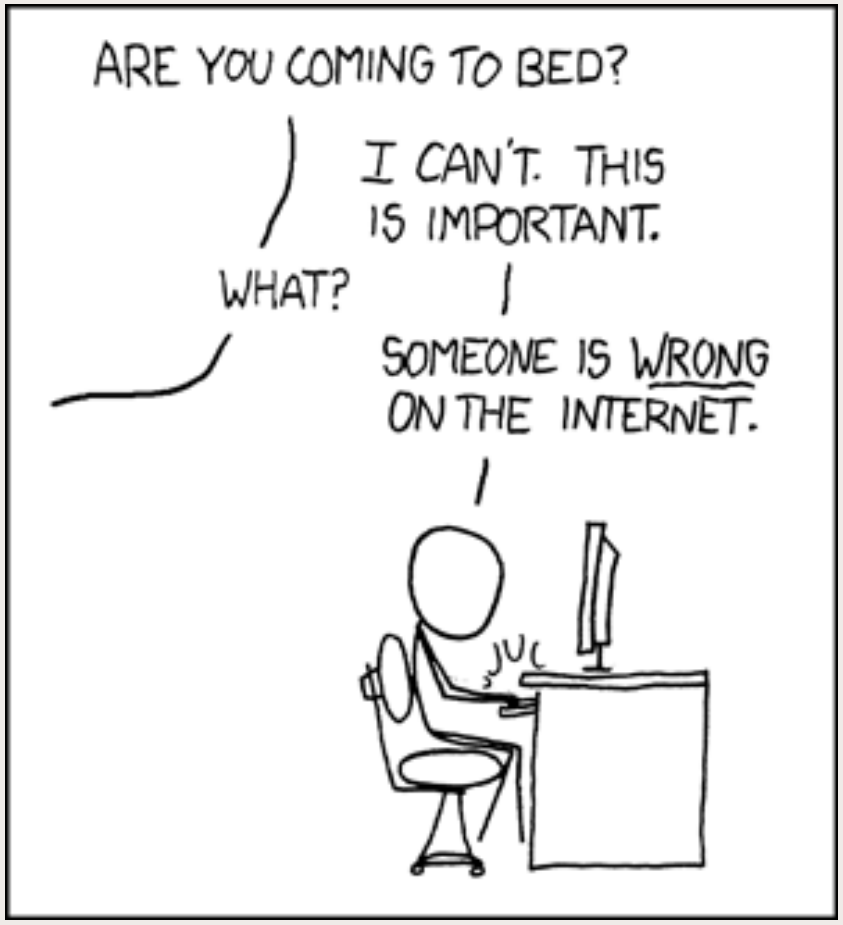
Benjamin Ooghe-Tabanou (@boogheta@piaille.fr / @boogheta)
Sciences Po médialab (@medialab_ScPo)

SciencesPo
MÉDIALAB

Exploiter le Web comme terrain d'enquêtes

Le Web : un espace de débats et de controverses

Collecter, enrichir, nettoyer, visualiser & analyser les traces numériques



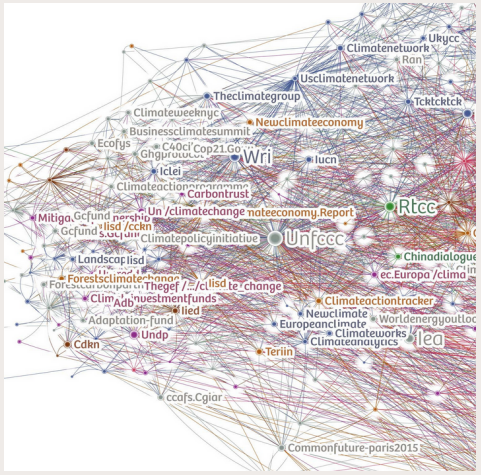
SciencesPo
MÉDIALAB

Hyphe : un crawler orienté recherche

http://hyphe.medialab.sciences-po.fr/demo/

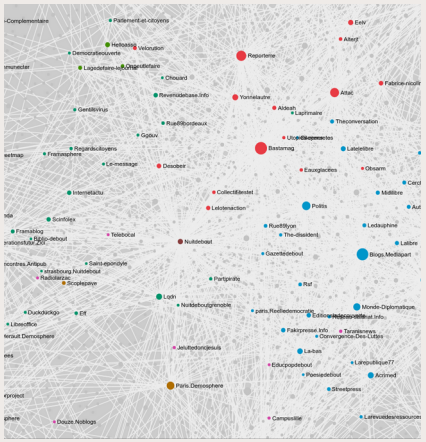
- Les liens hypertextes : nouveaux révélateurs de relations entre acteurs d'une thématique
- Créer un corpus documentaire
  - liste d' « acteurs web » & contenus textuels respectifs
  - réseau des liens hypertextes entre ces acteurs
- Études exploratoires ou de controverses dans tout domaine

hyphe



http://medialab.github.io/double-dating-data/

COP 21
Vie privée
Extrême droite
Tissu associatif
Produits laitiers
Cellules souches
Administrations culturelles
…

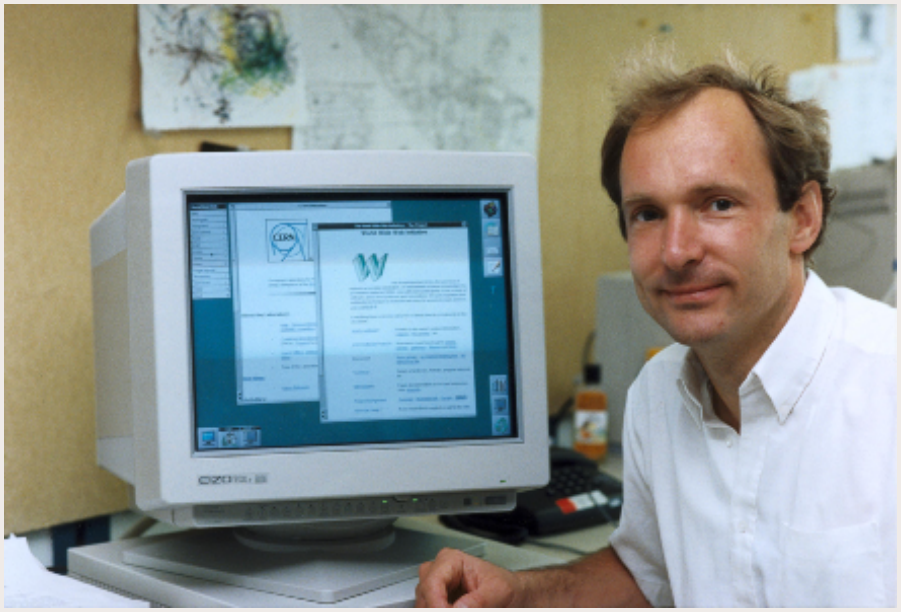


http://utopies-concretes.org/

OOGHE-TABANOU, Benjamin, JACOMY, Mathieu, GIRARD, Paul & PLIQUE, Guillaume, "**Hyperlink is not dead!**", In Proceedings of the 2nd International Conference on Web Studies (WS.2 2018). ACM, New York, NY, USA, 12-18. DOI: https://doi.org/10.1145/3240431.3240434

Crawler le web en SHS : pour quoi faire ?

L' « **Hyperlien** » au cœur de la conception du Web
→ porteur de sens et de structure



« *The texts are **linked together** in a way that one can go from one concept to another to find the information one wants.*

*The network of links is called a **web**. […]*

*The texts are known as **nodes**.*

*The process of proceeding from node to node is called **navigation**.* »

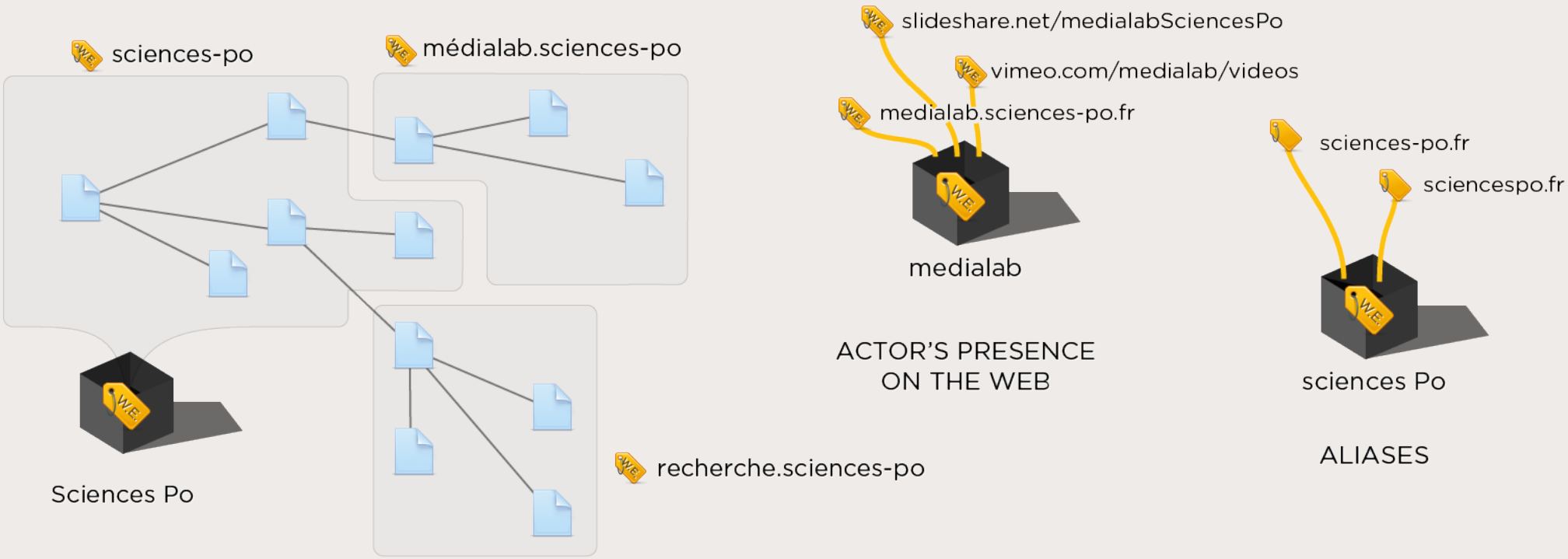
Tim Berners-Lee, 1990, WorldWideWeb: Proposal for a HyperText Project

« *A hyperlink is a **manifestation of intention**. **By linking** one page to another, one piece of text to another, **people intend** to do particular things.* »

Ryfe, Mensing, & Kelley, 2016, What is the meaning of a news link?

Définir finement les frontières de ses acteurs

Mais qu'est-ce donc qu'un « site web » ?

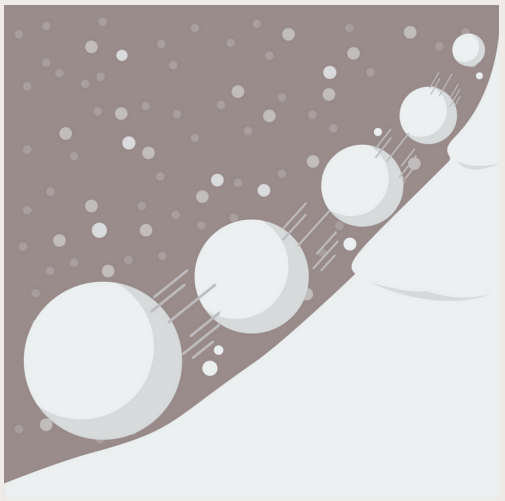


→ « **WebEntité** » : ensemble de pages web agrégées pour rassembler l'incarnation précise d'un acteur sur le web au sens d'une question de recherche spécifique

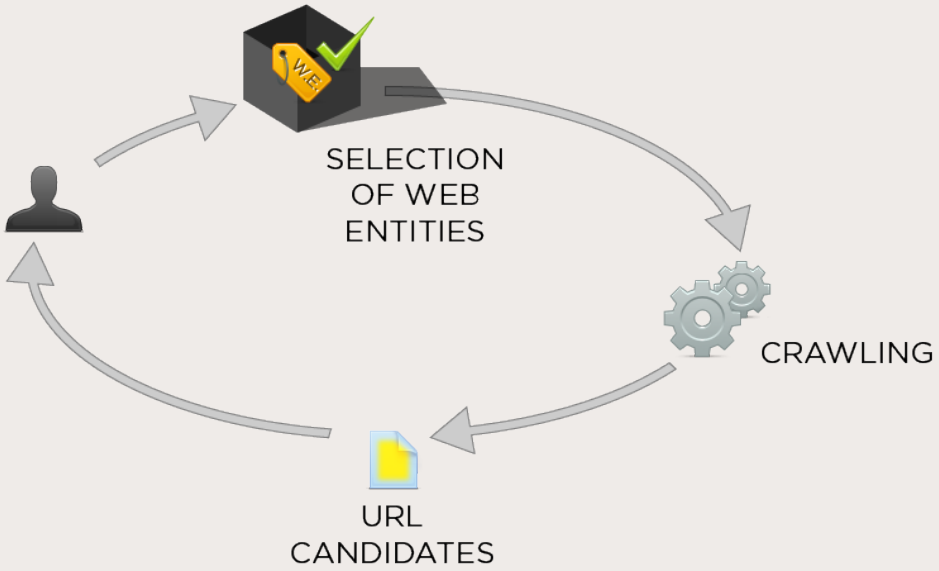
=> ensemble de préfixes d'URLs

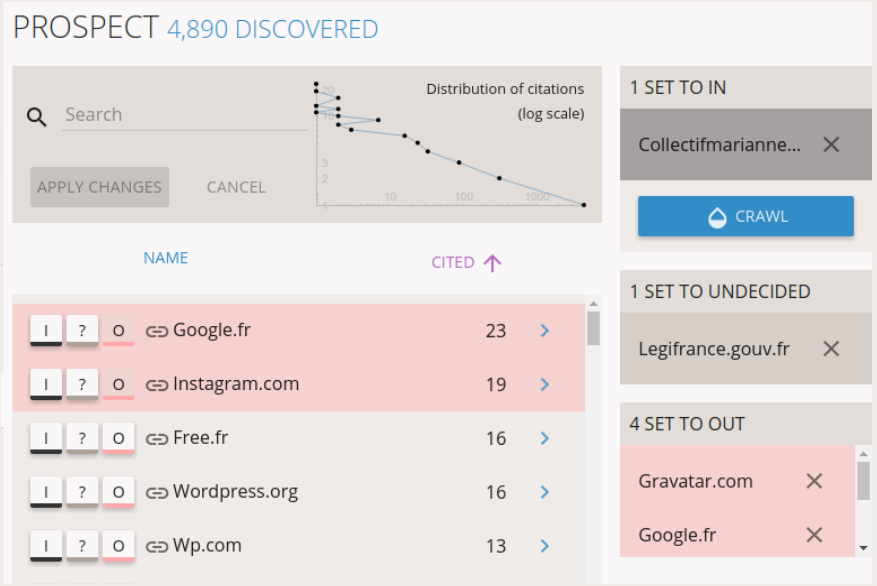
Une stratégie de crawling contrôlée

- Crawlers classiques : snowballing
  → Surreprésentation des couches hautes (Google, YouTube, Wikipedia…)
  → Dérive thématique rapide



- Hyphe : crawling semi-automatique
  - Fouille systématique uniquement des pages des WebEntités choisies
  - Choix humain des autres WebEntités à crawler grâce au degré de citation



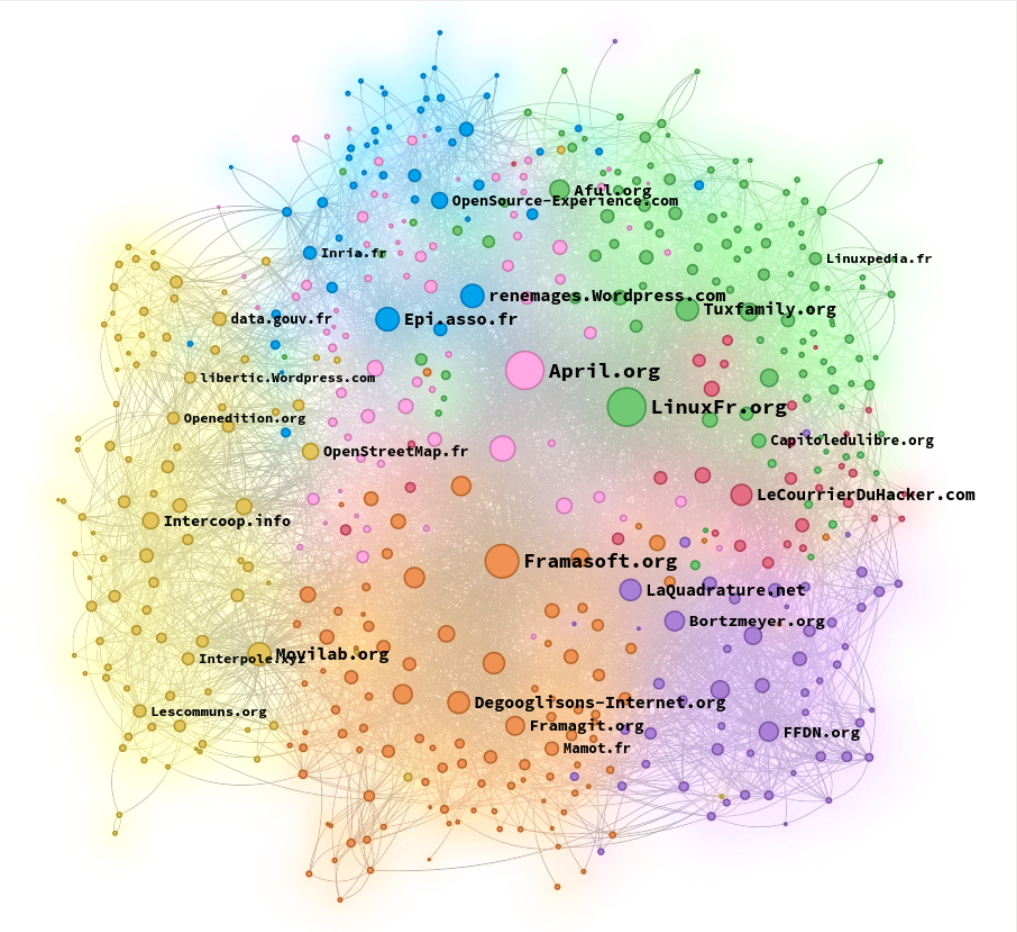


De multiples usages pour Hyphe

- Une méthodologie complète :

  sourcing, curation semi-automatisée, construction itérative, analyse exploratoire, catégorisation qualitative, visualisation de réseaux, analyse quantitative

- Divers publics-cibles :
  - Recherche : équiper les chercheurs en SHS pour réaliser un terrain web
  - Pédagogie : enseigner aux élèves le web au-delà de Google & Facebook
- Des analyses locales ou globales :
  - la structure interne d'un site web
  - Cartographie des communautés web sur une thématique
  - alliances et oppositions entre acteurs d'une controverse
  - etc.



Communautés du Libre en France sur le web

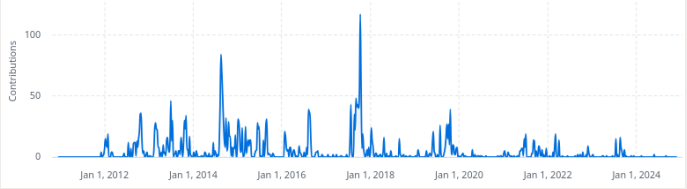
Le développement d'outils au médialab

https://medialab.sciencespo.fr/outils/

- Viser une large **Adoption :**
  - → **conception** d'outils dédiés aux besoins des chercheurs
  - → **design** d'interfaces centrées sur l'utilisateur
  - → **publication** d'outils web utilisables directement en ligne

- Assurer un maximum de **Réutilisabilité :**
  - → diffusion **au plus tôt** en **Logiciel Libre**
    (téléchargeable, installable, vérifiable & modifiable)
  - → développement « **opportuniste** » et « **agile** »
    « Release early, release often »
  - → mise à disposition d'**APIs**

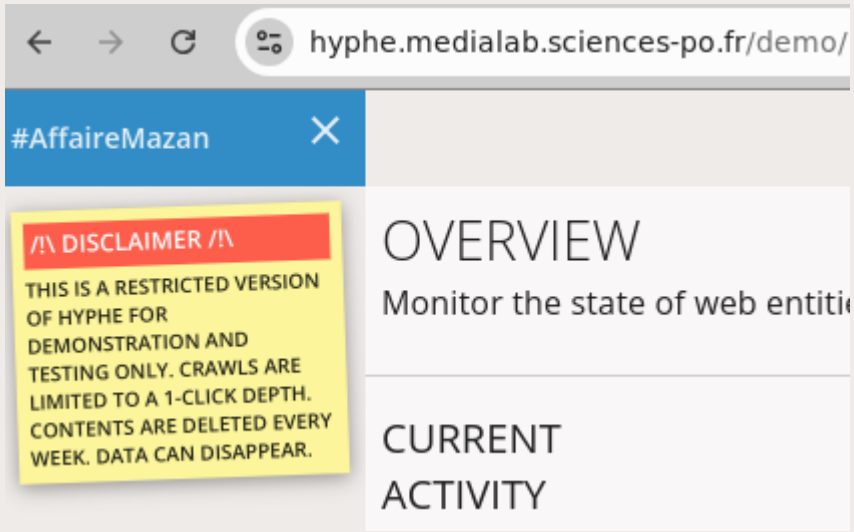
- **Documentation académique** et pratique
  (publications scientifiques, tutoriels, formations...)



v1.11.0 on Dec 14, 2023	v1.10.6 on Jul 21, 2023	v1.0.3 on Aug 3, 2018	v0.2.2 on Jun 20, 2017
v1.11.0-alpha on Sep 8, 2023	v1.10.5 on Nov 29, 2022	v1.0.2 on Apr 26, 2018	v0.2.1 on Feb 12, 2016
v1.10.9 on Aug 25, 2023	v1.9.1 on Aug 19, 2022	v1.0.1 on Jan 26, 2018	v0.2 on Apr 23, 2015
v1.10.8 on Aug 21, 2023	v1.9.0 on Mar 11, 2022	v1.0.0 on Jan 15, 2018	v0.1 on Feb 7, 2014
v1.10.7 on Aug 3, 2023	v1.9.0-alpha3 on Feb 14, 2022	v0.3 on Oct 16, 2017	v0.0.0 on Jul 19, 2013

Faciliter l'adoption et l'installation

- Version de démonstration en ligne restreinte



~ 500 corpus
de test créés
chaque année

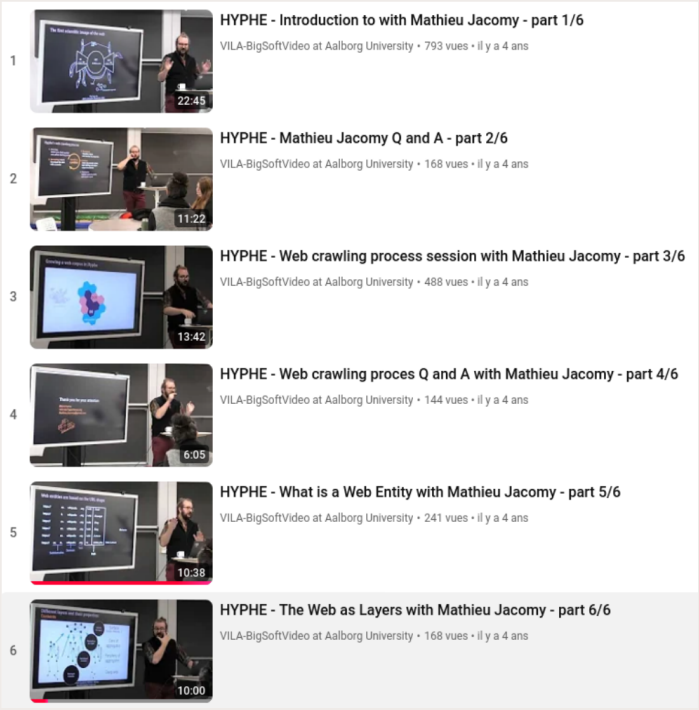
- Simplifier l'installation multi-plateformes
  → Linux, Mac & Windows via Docker !

- Accompagner les installations individuelles et institutionnelles
  → guider et automatiser des installations cloud

- Proposer un service payant d'installations SAS

Accompagner les nouveaux utilisateurs

- Proposer des tutoriels texte ou vidéo



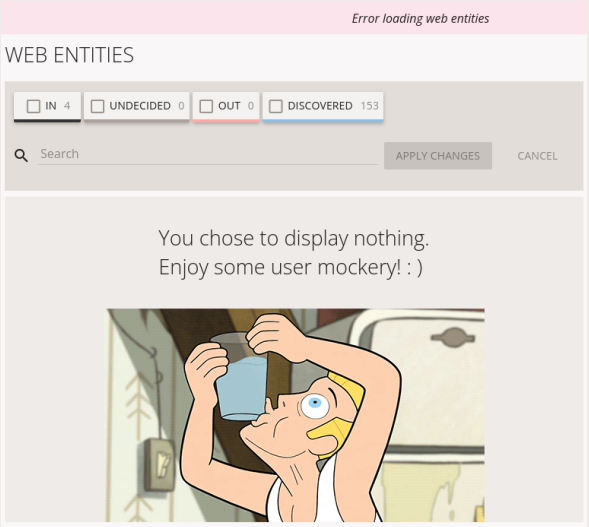
- Publiciser et expliquer les outils

- séminaires
- conférences
- formations
- cours de master
- ateliers pratiques
- écoles d'été/hiver
- ...

- Investir sur l'Expérience Utilisateur (UX)

  concevoir des interfaces tous publics

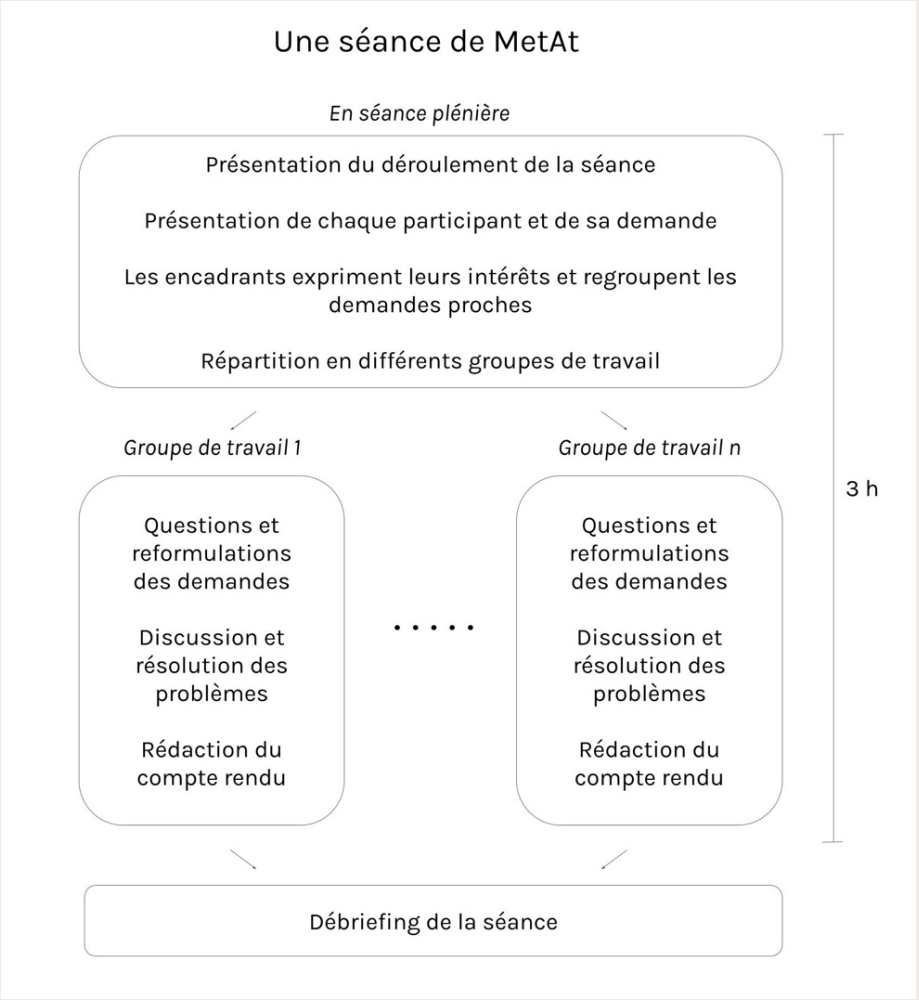
  → designer d'interfaces est un métier !



Le MetAt, atelier de méthodes numériques

https://medialab.sciencespo.fr/activites/metat-atelier-de-methodes/

- Demandes d'accompagnement :
  - discussion et conseil méthodologique
  - formation aux outils
  - collecte & nettoyage de données
  - visualisation exploratoire
  - …
- Un mardi après-midi par mois
- Ouvert à tous, sur inscription préalable
- Initialement « atelier du médialab » :
  → canaliser les sollicitations
- Élargi à la communauté des ingénieurs de recherche de Sciences Po en 2017 et à ceux de l'UPC en cours
- Contribue à l'autoformation continue

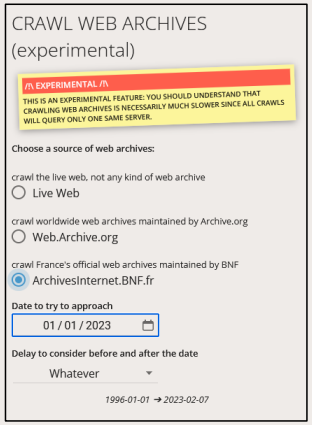


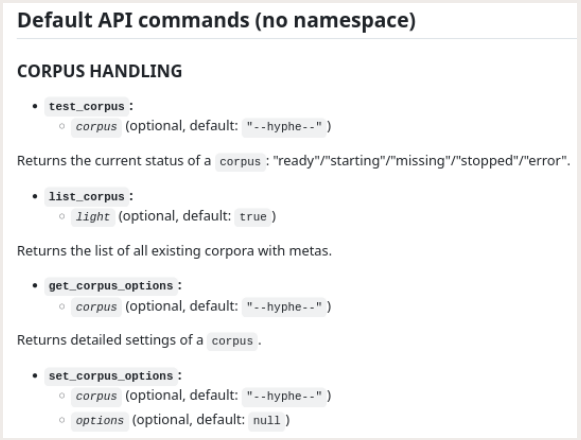
Diego Antolinos-Basso, Audrey Baneyx, Héloïse Théro, Benjamin Ooghe-Tabanou and Paul Girard, **"L'atelier de méthodes de Sciences Po : apprendre, aider, rassembler"**, Humanités numériques, 5 | 2022, Online since 01 June 2022, connection on 09 November 2022.
http://journals.openedition.org/revuehn/2799 DOI: https://doi.org/10.4000/revuehn.2799

Favoriser l'émergence d'un écosystème d'outils

- Mettre en place des partenariats (BnF, INA, etc.)
- Proposer une API exploitable par d'autres outils

  → interne : Hyphe-Browser, minet …
  → externe : IssueCrawler, PandorÆ …





Default API commands (no namespace)

CORPUS HANDLING

- `test_corpus` :
  - *corpus* (optional, default: `"--hyphe--"`)

Returns the current status of a `corpus` : "ready"/"starting"/"missing"/"stopped"/"error".

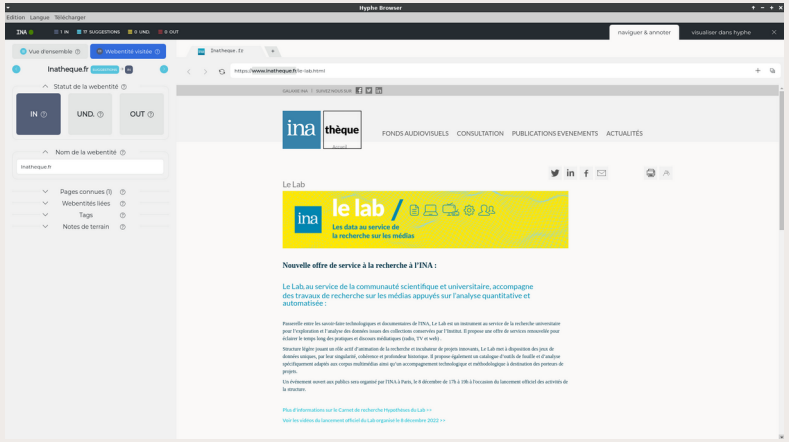
- `list_corpus` :
  - *light* (optional, default: `true`)

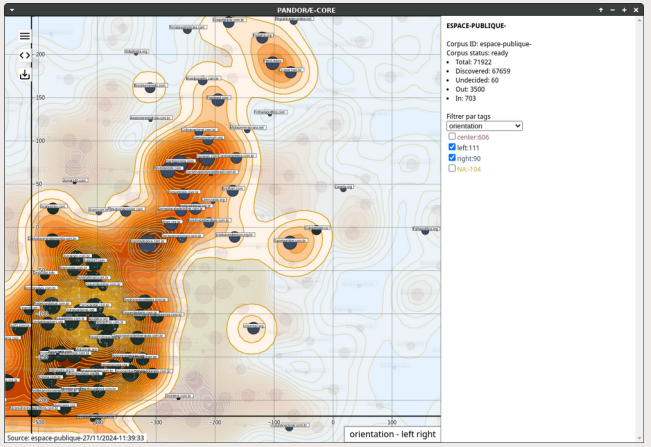
Returns the list of all existing corpora with metas.

- `get_corpus_options` :
  - *corpus* (optional, default: `"--hyphe--"`)

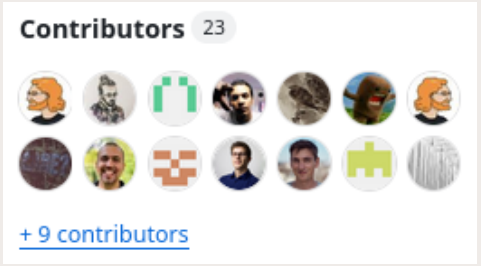
Returns detailed settings of a `corpus`.

- `set_corpus_options` :
  - *corpus* (optional, default: `"--hyphe--"`)
  - *options* (optional, default: `null`)





- Proposer une documentation destinée aux développeurs (architecture, code, installation, configuration…)
- Encourager (et accepter !) les Pull Requests

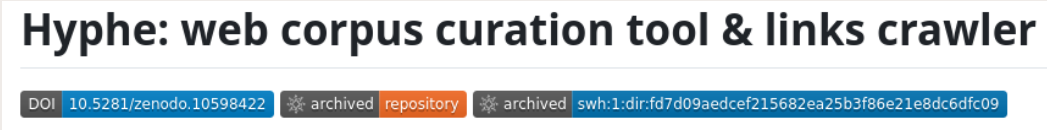


Documenter et valoriser scientifiquement

- Permettre la **citation académique** de l'outil :
  - DOI pour le logiciel (via Zenodo)
  - Archivage via Software Heritage
  - Publications méthodologiques peer-reviewed
    - Jacomy M., Girard P., Ooghe-Tabanou B., Venturini T. (2016), **Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences**, ICWSM 2016, Cologne

      https://sciencespo.hal.science/hal-01293078v2
    - Ooghe-Tabanou B., Girard P., Jacomy M., Plique G. (2018), **Hyperlink is not dead!,** ACM Proceedings of the 2nd International Conference on Web Studies (WS.2 2018) Paris.

      https://dl.acm.org/doi/10.1145/3240431.3240434



- Publier également auprès des communautés logicielles :
  - Présenter les challenges techniques (devroom Open Research @ FOSDEM)
    - Plique G., Jacomy M., Ooghe-Tabanou B., Girard P. (2018), **It's a Tree... It's a Graph... It's a Traph! Designing an on-file multi-level graph index for the Hyphe web crawler**, FOSDEM 2018, Bruxelles

      https://medialab.github.io/hyphe-traph/fosdem2018/#/

Documenter et valoriser scientifiquement

Title	Creator	Publication Title	Date
Online networking behaviour of tourism stakeholders in a multi-desti…	Herasimovich et al.	Journal of Destination Marketing & M…	2024-03-01
Varying Roles of Destination Management Organizations in the Digit…	Herasimovich et al.	Information and Communication Tech…	2024
Green politics beyond the state: radicalizing the democratic potential…	Ejsing et al.	Climatic Change	2023-05-30
Gruson-Daniel, Célya. “Mapping Contemporary “research on Resea…	Community	Open Research Community	2023-03-15
Mapping European Digital Heritage Politics: An Empirical Study of E…	Capurro and Severo	Heritage & Society	2023
InSciC—Knowledge-Aware Crawler for Indian Sciences	Hegade et al.	Proceedings of International Confere…	2023
The ‘more-than-food’ geographies of omega-3s	Browne		2022-10
Atlas multi-plateforme d’un mouvement social : le cas des Gilets jau…	Morales et al.	Statistique et Société	2022-09-28
When teachers Google physical literacy: A cartography of controver…	Young et al.	European Physical Education Review	2022-08-01
A network view on reliability: using machine learning to understand …	Blanke and Venturini	Journal of Computational Social Scie…	2022-05-01
Research Methodologies and Ethical Challenges in Digital Migration…	Sandberg et al.		2022
Controversy Mapping and the Care for Climate Commons	Papazu and Veng		2022
Réinterroger les notions d'accès à l'information géographique numér…	Desbonnet		2021-09-01
Making the circular economy online: a hyperlink analysis of the artic…	Humalisto et al.	Environmental Politics	2021-07-29
Savoirs incertains	Boullier et al.	RESET. Recherches en sciences soci…	2021-05-20
Uncertain Knowledge. Studying “Truth” and “Conspiracies” in the Di…	Boullier et al.	RESET. Recherches en sciences soci…	2021-05-20
“I’m not an antivaxxer, but…”: Spurious and authentic diversity amo…	Cafiero et al.	Social Networks	2021-05-01
Communication network analysis to advance mapping ‘sport for dev…	Herasimovich and Alzu…	International Review for the Sociology…	2021-03-01
Mapping the spread of Russian and Chinese contents on the French…	Douzet et al.	Journal of Cyber Policy	2021-01-02
Beyond collected data: Politics of APIs on social media platformsa			2021
Étude exploratoire sur la « recherche sur la recherche » : acteurs et …	Gruson-Daniel and And…		2021
Machine Learning based Classification of Online News Data for Dis…	Gopal et al.	2020 IEEE Global Humanitarian Tech…	2020-10
La vulgarisation des recherches sur le phénomène de harcèlement …	Stassin	Revue française des sciences de l’inf…	2020-09-01
Ville intelligente et e-gouvernance en Inde, cartographier un nouvea…	Leclerc	Mappemonde. Revue trimestrielle sur…	2020-07-01
Exploratory analysis of the hypertext structure linked to diabetes	Shi		2020-06-23
Cartographier la propagation des contenus russes et chinois sur le …	Douzet et al.	Herodote	2020-06-17
Communication network analysis to advance mapping ‘sport for dev…	Herasimovich and Alzu…	International Review for the Sociology…	2020-03-15
Beyond issue publics? Curating a corpus of generic Danish debate i…	Munk and Olesen	STS Encounters	2020-02-08
A New Web-Based Big Data Analytics for Dynamic Public Opinion M…	Tournay et al.	OMICS: A Journal of Integrative Biology	2020-01

Merci de votre attention !

Code source de Hyphe :

https://github.com/medialab/hyphe

Version de démo limitée en ligne :

https://hyphe.medialab.sciences-po.fr/demo/

benjamin.ooghe@sciencespo.fr

@boogheta@piaille.fr @boogheta @medialab_ScPo