Exploring web archives using Hyphe, a research oriented web crawler

Sara Aubry, BnF & Benjamin Ooghe-Tabanou, médialab Sciences Po
IIPC Web Archiving Conference, April 25-26th 2024, Paris

Connecting Hyphe to web archives, an outcome of the project

Methodological principles:
• Web entities: fine definition of actors beyond "websites"
• Prospection loop: iterative curation of discovered actors

National Library of France web archives
• From 1996 to now
• 52 billion URLs, 2 PB of data
• Growing with annual crawls of million websites (5.6 mio in 2023) and regular crawls of thematic, event and media focused websites (currently 70,000)

Connecting Hyphe to web archives, an outcome of the project

Hyphe is now compatible with web archives from BnF, INA, Arquivo.pt & Internet Archive

Experiment crawling through the archives in a DataSprint

• Hands on workshop during 5 days within BnF DataLab
• Dive into data to explore hypothesis on research questions
• Gather complementary skills & expertises: research, web archiving, engineering, design, digital methods
• Experiment with different methodological and exploratory approaches
• Find preliminary insights

Focus on a presidential candidate’s campaigns: Jean-Luc Mélenchon

Hypothesis: online communities changed across 2012, 2017 & 2022 campaigns

Methodological approach:
• Define time windows depending on available archives
• Start coherent crawls from stable sources (Wikipedia)
• Define common rules to select and include actors
• Use ontologies to tag actors (editorial form, nature)
• Produce graphs to visualize differences
• Identify disappeared actors, newcomers and resilient ones

- Hyphe’s source code, demo & references: https://github.com/medialab/hyphe
- BnF’s Web Archives introduction: https://www.bnf.fr/fr/depot-legal-du-web
- ResPaDon DataSprint results: https://respadon.medialab.sciencespo.fr
- Contacts: sara.aubry@bnf.fr, benjamin.ooghe@sciencespo.fr
- Photo credits: Caroline Maufroy / Sciences Po