



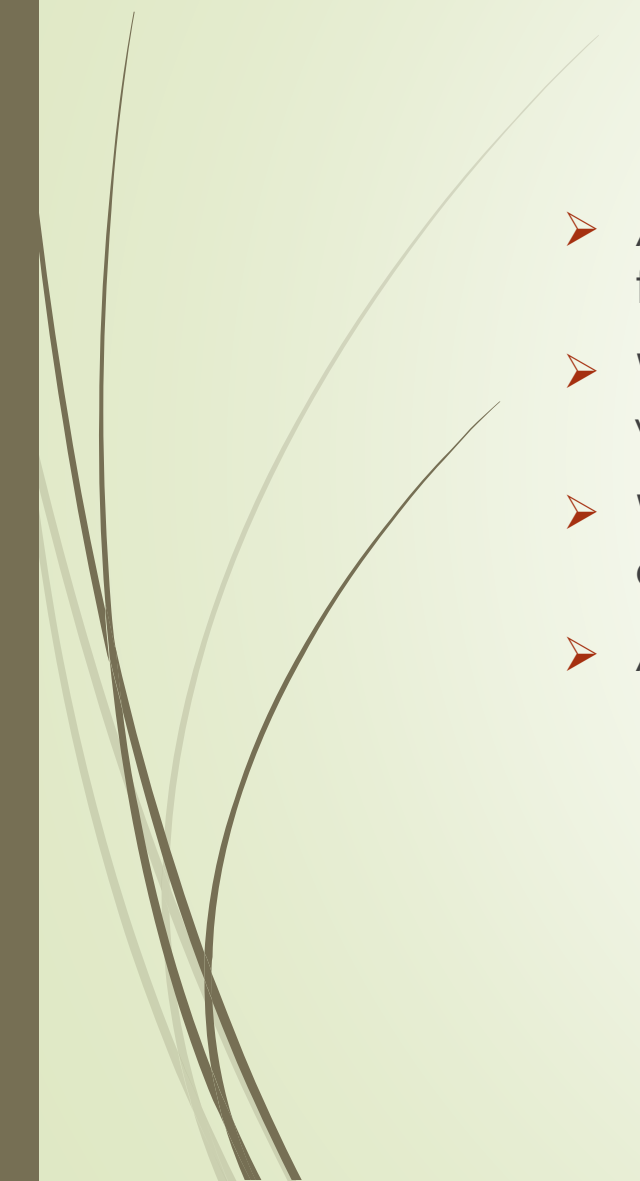
# How can we build a corpus and an archive of the French web literature?

Christian COTE

MARGE/Université Lyon3



# What is web literature?

- A literature of anonymous people (amateur writers, preferring short texts, frequent publications).
  - Writers identifiable by their mutual recognition (few isolated writers: role of writing workshops, associations, meetings...).
  - Writings, images and comments: use of the properties of multimedia. (For example: capacity of the images to elicit comments: not fixed sense).
  - Amateur and massive practice.
- 

# Overview of the LIFRANUM Project

Objective: a platform of web literary production scientific resources: a corpus and an archive.

## Challenges for partners:

MARGE: digital literature: visibility and collection

ERIC: computer science (data lake, NLP): structuring and indexing

BNF: digital literature and web archives: developing the specialized web archive.

- Duration: 4 years
- Collect/record phase involving BnF in the partnership: **development of a workflow** (double crawl).
- Conservation perspective (BnF's mission) AND corpus available to researchers.
- New modes of research. **Analysis of literary productions thanks to machine learning:** to make writing proximities.
- Methodology: **a workflow based on method for identifying, collecting and structuring** the identified resources to inform a crawl and finally an evaluation, linked to the production of the archive (and carried out by the BNF).

# Workflow of the project

- The resource production chain: a succession and enrichment between two data sets [corpus and archive].

**The corpus: a structured object of study and a space for exploration.**

- We acquire URLs by prior identification in order to realize a first crawl.

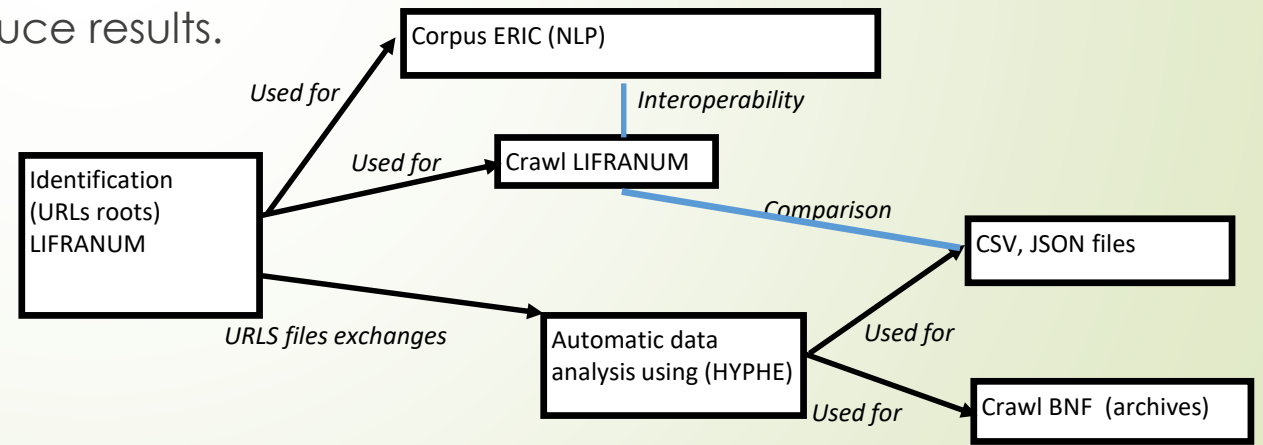
**The archive: a heritage of this literature in its social context.**

- We acquire complementary data by an automatic identification of the links associated with the previously identified URLs -> constitution of the archive by a second crawl.
- The results of the second analysis thus complete those of the first, by integrating URLs outside the previously identified writers' networks.

	Manual procedure	Automatic procedures	Manual procedure
LIFRANUM Corpus	Identifying resources of interest and XML directories	HERITRIX crawling without control of the process flow	
BNF Archive		Link analysis and supervised crawling	Link tracking using HYPHE

# Working tools and methods

- Addition of a working corpus (ERIC laboratory) to experiment indexation.
- Working materials: JSON, CSV and EXCEL files to exploit the data obtained automatically (HYPHE, BERT for textual analysis).
- Principles of situated and distributed cognition: human operations are systematically recorded and filed in a way that can be reused.
- Tools produce data but not results.
- Human evaluation of data to produce results.






# Corpus, archives: definitions

Different uses between the corpus as "primary data of science" and the archive as heritage accessible on site.

Experience of corpus linguistics applied to the web: *the web as an infinite repository of written linguistic achievements. Any production is by definition relevant.*

- Capacity of a large quantity of data to observe regular phenomena behind surface variations.
- Detection of evolutions and identification of regularities that only appear with massive data and quantitative tools.
- Choice of a massive crawl in order to have a sum of data that can be explored through a unique portal.
- About 1000 writers (before crawl), most of them unknown.
- Indexing based on : types of discourse (narrative, descriptive, etc.), displayed identity of discourse (psychological, meta-discursive, communicational, etc.), aesthetic preference.





# Fundational questions about the LIFRANUM identification.

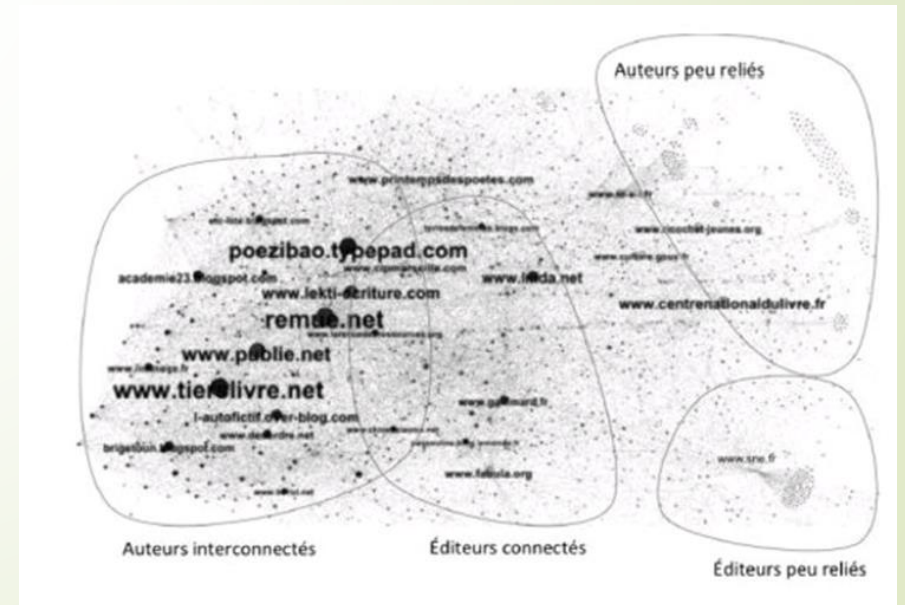
How can we identify web literary production in all its diversity?

- No key words or specific or recurrent indices (unlike COVID archives). Hence the **sociability networks** are an entry point to the web.
- Social networks are both a **social and technological fact**.
- **Identification: an information chain to collect and structure URLs to constitute a directory for the crawl.**
- 1 engineer over one year.

# Networks and recognition in web literature

Foundation of networks: **peer recognition as a marker of literarity**. (A textual production becomes literary within the framework of the Web by the recognition of the peers, marked by links of the type "sharing of URLs").

- **Many "relays"**: digital works taken up by others. Dispersion of the works and productions of authors (in different places of publication).
- **But communities of practice from which the other spheres are organized**. Importance of markers and digital spaces of sociability (communities).
- Inadequate author/work indexing unit.





# Protocol for relevant URLs identification

GOOGLE Structured Advanced Search: only relevant and noise-free URLs.

Formal framework of situation theory:

- Distinction between a **world individual (the named entity)**, the **mental universes (the property)** and the **context** in which this relationship operates (the web publication media).

By placing the individual at the center of the query, we limit the results to those that concern him. Each URL pointing to this author can have as its author a candidate for a new query.

The image shows a screenshot of the Google Structured Advanced Search interface with three annotations. The first annotation points to the 'tous les mots suivants' field containing 'Ecrivain, écrit'. The second annotation points to the 'ce mot ou cette expression exact(e)' field containing 'Makenzy Orcel'. The third annotation points to the 'l'un des mots suivants' field containing 'site blog page Facebook Haïti'.

Trouvez des pages avec...	
tous les mots suivants :	Ecrivain, écrit
ce mot ou cette expression exact(e) :	Makenzy Orcel
l'un des mots suivants :	site blog page Facebook Haïti
aucun des mots suivants :	

Property of a person considered as writer and producer of literary contents

Exact individual in the world

Object name of the web platform of edition or editor (and geographic name if required)

# Typing the URLs

- Query results reported in XML schemas. Each XML schema = 1 result (all the links that appeared).
  - Ingoing and outgoing links entail mutual recognition.
  - Links and forms of recognition differ according to the literary project: typing of projects by kinds of links exchanged.
- **Individual authors:** creative production signed by a person name.
  - **Collective authors:** different people identified with a unique signature.
  - **Communities:** name associated to a place and initiatives linking different individual and/or collective authors.
  - **Supports:** edition of original creations of individual or collective authors.

## Categories of literary projects

Individual author

Collective author

Writers community

Support (periodic)

# Networks record

Structuring of the diagrams by specifying these different forms of links:

- Link 0: **outgoing links of the page** referring to other achievements of the author's project.
- Link 1: **outgoing links of the page (direct link)** that indicate a relationship with another literary project
- Link 2: **incoming links highlighted by the information search** (direct quote). It is the URL found that points to the original search. (inverse relationship of link 1)
- Link 3: **links highlighted by the information search** (indirect citation).

The identified URL refers to the search object but without a URL link.

Empirical but potentially reproducible modeling : information flow.

## Characterization of links

Links of origin of the author's mention or reference

Outgoing links from the author's site

Inbound links by referencing the author's site on another site

Links by quotation and mention (without referencing)

```
<schema elementFormDefault="qualified" targetNamespace="http://www.example.org/LIFRANUMidentification">
  <element name="collection" type="string"/>
  <complexType name="network">
    <attribute name="description">
      <simpleType>
        <restriction base="string">
          <enumeration value="personal/communaury"/>
          <enumeration value="personal/personal"/>
          <enumeration value="communaury/communaury"/>
          <enumeration value="communaury/personnel"/>
          <minLength value="0"/>
          <maxLength value="1"/>
          <enumeration value="value"/>
        </restriction>
      </simpleType>
    </attribute>
  </complexType>
  <complexType name="facet">
    <complexContent>
      <extension base="tns:network">
        <sequence>
          <element name="webunit" type="string"/>
        </sequence>
        <attribute name="provenance" type="string"/>
        <attribute name="link0" type="string"/>
        <attribute name="link1" type="string"/>
        <attribute name="link2" type="string"/>
        <attribute name="link3" type="string"/>
        <attribute name="authorproject" type="hexBinary"/>
        <attribute name="communityproject" type="hexBinary"/>
      </extension>
    </complexContent>
  </complexType>
</schema>
```

# Crawl and addresses sharing

- **A strongly constrained crawl** (limit out-of-field branches). Crawl by host or by domain (for blogs depending on a particular API).
- Limit of **two hops** from the root. Maximum **depth of** crawls: restore all work.
- Web literature strongly inscribed in a context built by the author (individual or collective).
- HERITRIX for the crawl and SOLRWAYBACK/AUT as exploration tool.

	Manual procedure	Automatic procedures	Manual procedure
LIFRANUM Corpus	Identifying resources of interest and XML directories	HERITRIX crawling without control of the process flow	
BNF Archive		Link analysis and supervised crawling	Link tracking using HYPHE



# Inverse methodology and verification

- Question of granularity: WARCs = URLs, not hosts or roots.
- XML files are anchored to crawled URL roots.
- Manual method of identification: limit a field.
- But GOOGLE => bias.

An automatic method completes the first one: **authors indirectly associated with the first ones.**

The root URLs are reinterpreted by automatic link analysis using HYPHE:

Other addresses that would not be recognition links: authors that **gravitate around the constituted networks.**

Duality between the corpus and the archive: **distinction between literary projects and literary domain.**



# Crawls issues

Number of URLs in the starting list of HYPHE -> small sample (WORDPRESS accounts).  
HYPHE analysis: impressive set of links -> manual selection (between 10 and 20% of relevant sites). 10 months of engineering.

Distinction between :

- **Recognition links** between writers forming a network (initial identification)
- **Relationships to these networks** discovered by hypertext links alone

Complementarity between the two phases of the methodology: the second improve the results of the first.

LIFRANUM repertory before crawl)	BNF BC web repertory	Discovered by HYPHE
937 URLs	152 URLs	646 URLs

URLs identified and crawled by HERITRIX	URLs only identified by HERITRIX (but not crawled)	URLs discovered by HYPHE and relevant for the LIFRANUM corpus	URLs not relevant for the LIFRANUM corpus
2	46	100	498



# Result (before indexation)

We propose three different types of objects:

- A research corpus corresponding to the LIFRANUM crawl (+ materials: identification XML schemas, HYPHE visualizations, comparison data): scientific data to be exploited.
- A specialized replayable domain archive, 1TB -> no images and depth.
- A working corpus (TAL): auxiliary textual corpus, constituted from WORDPRESS and BLOGGER APIs (motivated by difficulties to work directly with the results of the crawl ("clean" data for an automatic analysis)).



# Indexation strategy

- Usual categories of literary classification (novel, poetry, short story, etc.) not relevant.
- Researchers' practices: navigation by **proximity**: recommendations by common features (of types of discourse, claimed scope and forms) or breaks between texts.

Indexing without prior controlled language.

Two complementary strategies:

- A **supervised strategy**: grammatical forms (e.g. argumentative connectors and verbs) -> discursive structures and discourse postures
- An **unsupervised strategy**: systematic vectorization of texts -> aesthetic choices (dynamics VS constants, diversity VS homogeneity)



# Conclusion

The work is still in progress, especially concerning a first indexation.

**Identification** : social networks, publication (WATTPAD) and forums : based on exchange modes other than URL sharing (+ impossibility for HERITRIX to crawl the contents of the platforms).

Links between the different forms of data and information representation:

- Two complementary **network analyses** (XML and HYPHE schemas),
- **indexing via a working corpus linked to the general corpus.**
- Finally, a tool to characterize **new literary forms and expressions, and systematicity allowed by the corpus.**



Acknowledgements :

- ▶ Lorraine Feugère, IGR MARGE
  - ▶ Kévin Locoh-Donou, IGR BNF
  
  - ▶ MARGE : Gilles Bonnet,
  - ▶ ERIC : Julien Velcin, Enzo Terreau, Javier Espinosa,
  - ▶ BNF : Alexandre Faye, Christine Genin,
- 