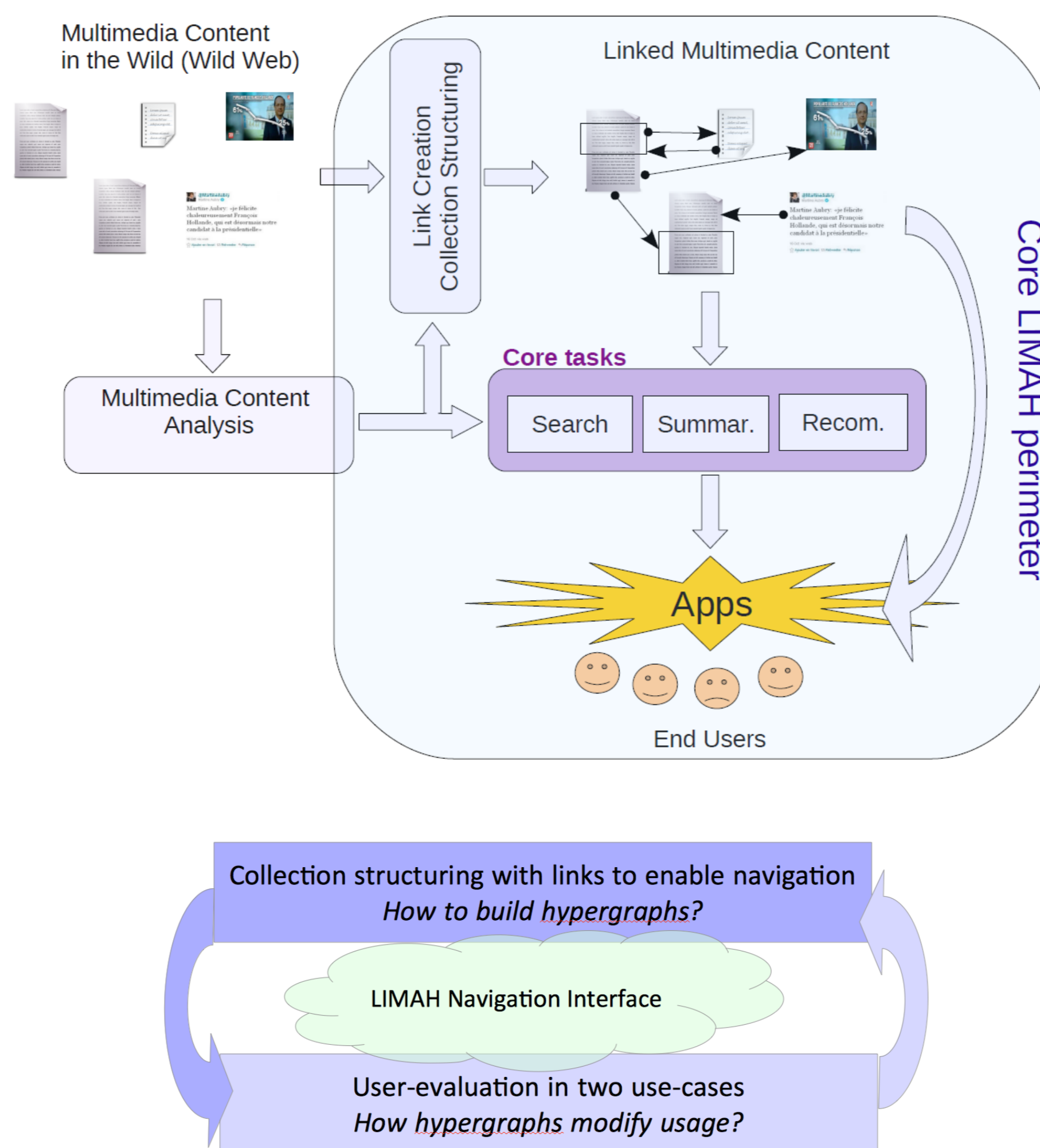


Overall description

Available multimedia content is rapidly increasing in scale and diversity, yet today, multimedia data remain mostly unconnected, i.e., with no explicit links between related fragments. The project Linking Media in Acceptable Hypergraphs (LIMAH) aims at **exploring hypergraph structures for multimedia collections**, instantiating actual links between fragments of multimedia documents, where links reflect particular content-based proximity—similar content, thematic proximity, opinion expressed, answer to a question, etc. Exploiting and developing further techniques targeting pairwise comparison of multimedia contents, LIMAH addresses two key issues of **content-based graph-oriented multimedia collection structuring**: How to automatically build from a collection of documents an hypergraph, i.e., graph combining edges of different natures, which provides exploitable links in selected use cases? How collections with explicit links modify usage of multimedia data in all aspects, from a technology point of view as well as from a user point of view? LIMAH will study hypergraph authoring and acceptability in **two distinct complementary use-cases**, namely, navigation in news data and learning with online courses.

In pictures...



Project organization

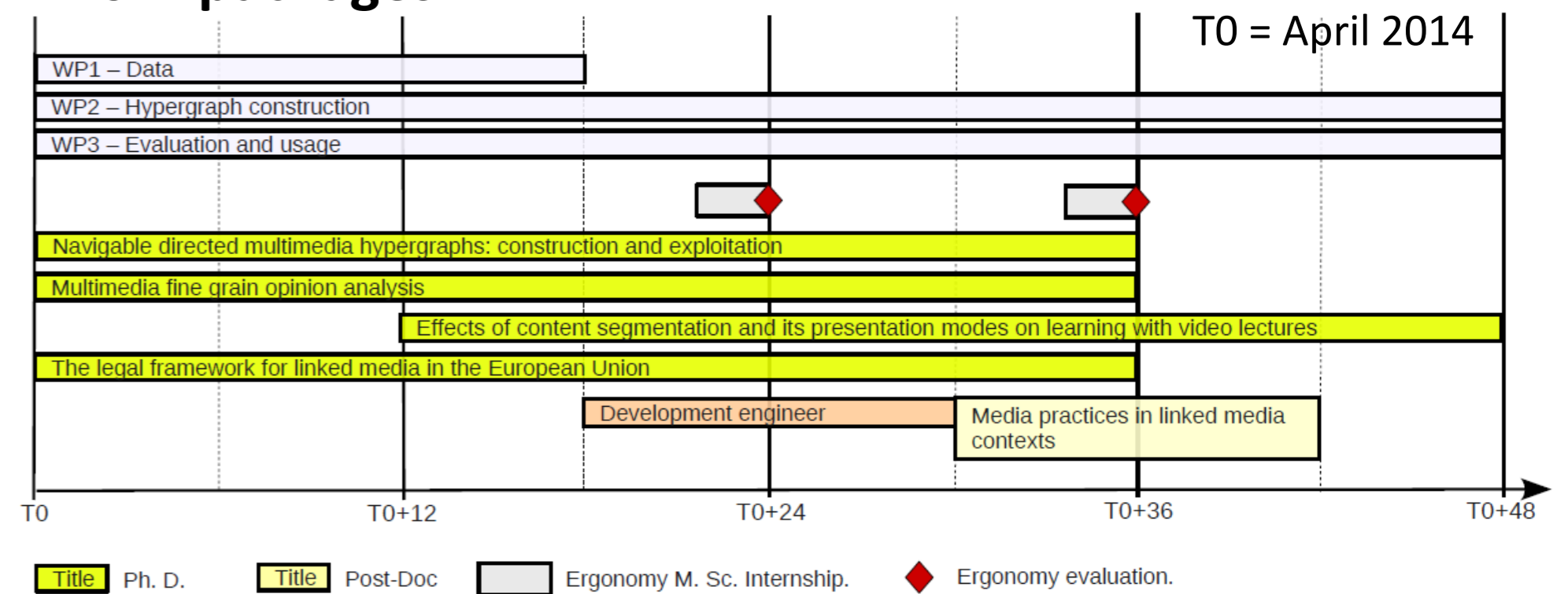
Consortium

LIMAH relies on a **pluridisciplinary consortium**, to apprehend linked media contents globally and develop a long term vision

- Technology partners
 - IRISA: Natural Language Processing
 - LINA: Natural Language Processing
- Usage and users
 - CRPCC: Psychology, Ergonomy
- Societal impact
 - IODE: Law and media
 - PREFics: Media sociology



Work packages



Use-case scenarios

Two application domains, **news and e-education**, with similar scenarios involving the exploration of a collection of (linked) multimedia content items.

Scenarios

1. Find a piece of information: *What is the Stroop effect? Who said "Moi Président"? When did we start talking about the Palestino-Israeli war?*
2. Global view on a subject, e.g., to make a synthesis: *start and end dates plus chronology, main facts, causes and consequences, people/organization involved, public opinion and main criticisms, etc.*

Semantically structured graphs

Linking multimedia fragments through the use of a directed hypergraph

- How to prune and/or combine hypergraphs to make them tractable?
- What data representation and comparison metric for what use?
- How to explain what links mean and bring to users?

Progress: Creation of a typology of links in news-related data; Topic-models for serendipitous content comparison

Legal issues

Legal feasibility, acceptability and legitimacy of linked media and related technologies

- How media linking technology could impact the field of Law by creating new practices and economic models?
- To what extent can legal sources in various fields (mainly copyright law, media law and fundamental rights) allow the development of linked media?
- What legislative changes and new legal solutions are required to permit the emergence of linked media services?

Progress: Identification of related legislation and jurisprudence; Categorization of issues and bibliography

Corpus

E-education data

- Limited number of courses
 - videos, textbooks, slides, Wikipedia, etc.
- Status: todo



News-related data

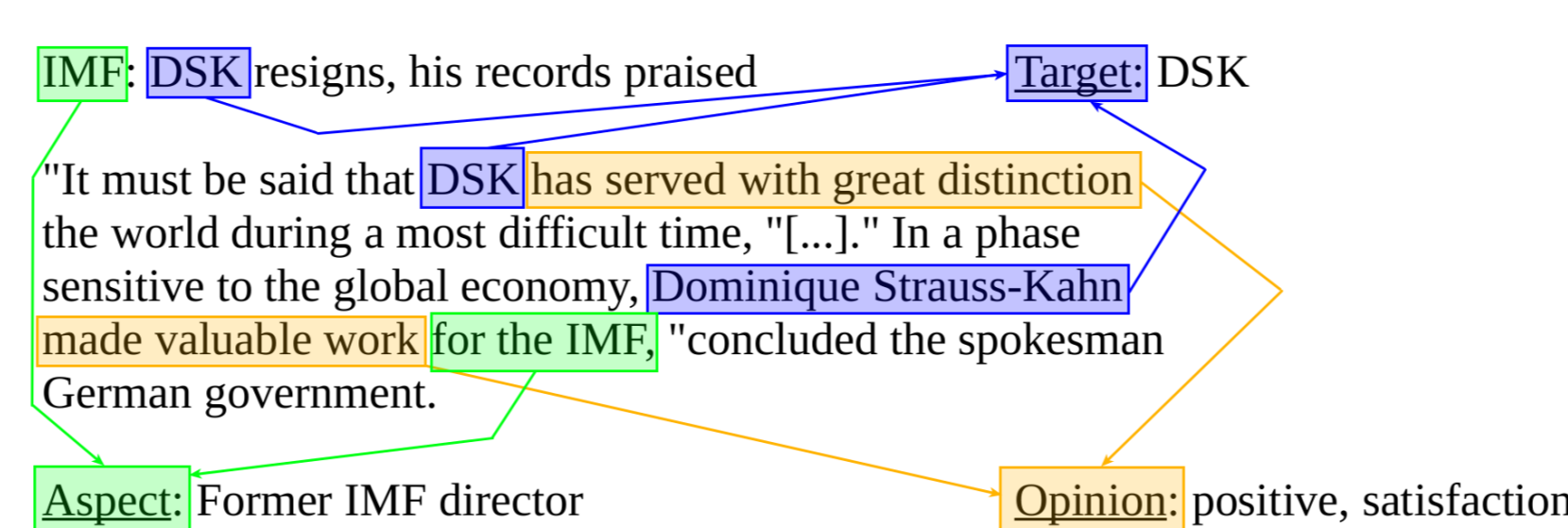
- Large amount of sources grabbed over 4 weeks
 - TV videos : news, debates, etc.
 - Online newspapers
 - Radio podcasts
 - Tweets and Facebook comments
- A few Gbytes per day
- A variety of content: multiple modalities, from daily 5 min. long news to monthly 2 h long debates
- Data + metadata + links in web pages
- Status: in progress



Opinion mining

Fine-grain opinion and sentiment analysis on news data to enrich content with opinion information

- How to deal with degraded data ""(ASR transcripts, tweets, Facebook posts)?
- How to precisely identify the target of the opinion and the different aspects in a collection?
- How to make use of (thread) context and account for temporal evolution of opinions?

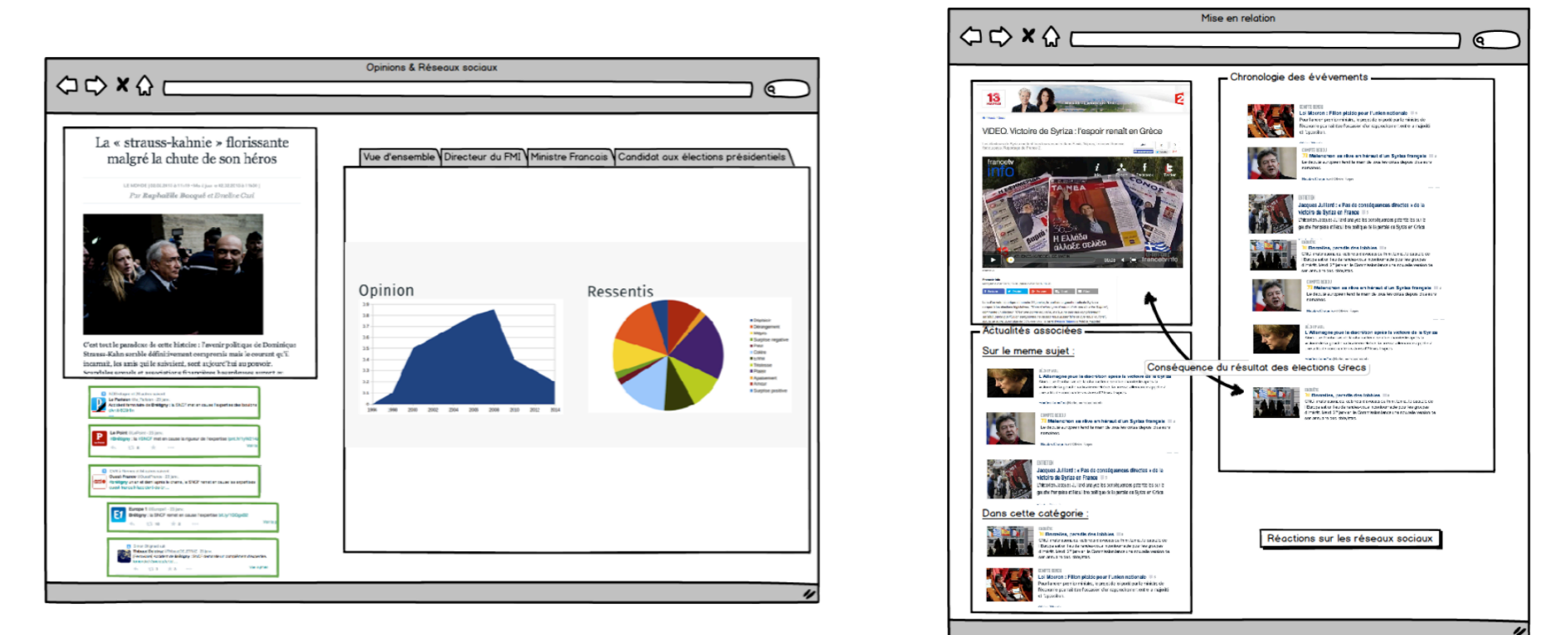


Progress: Creation of a pipeline to analyze opinions in tweets based on lexicons and unsupervised clusters

User expectations

Four categories of functionalities

1. Information highlight -> transcripts, keywords, named entities, links to external data (biographies, definitions, etc.)
2. Opinions and social networks -> opinions and feelings, globally or per aspect, temporal evolution, synthesis
3. Links and recommendation -> usefulness of various text distances, link typing and characterization
4. Rapid access and jump-in points -> table of content, segmentation, chronological organization, entity-based indexes, summaries



Methodology

1. Interviews based on description of functionalities and mockups: 5 to 10 interviews of 1h30 targeting journalists, press agents, students in Infocom
 - Understand practices and expectations
2. Online surveys on acceptability: ~100 participants
 - Define expected functionalities
3. Usability testing of the LIMAH navigation interface
 - Validate and improve functionalities

Preliminary trends from interviews

- High acceptance level by practitioners
- Variations with professional profile: importance of opinions and social networks for press agents, high added-value for transcription, rapid access and jump-in points for journalists
- Suggestion of new functionalities
 - Keyword ranking and organization
 - Smart filters for comments, keywords, etc.
 - Popularity-aware table of content
 - Automatic synthesis capabilities