IFLA-RBSC Mid-term conference

National Library of Norway

Oslo, April 12th 2018

EXPLORING THE DIGITAL HERITAGE OF READING: THE READ-IT PROJECT.

Anne Baillot[1], François Vignale[1,2]

[1]3L.AM, [2] Bibliothèques Universitaires
Avenue Olivier Messiaen, 72085 Le Mans cedex 9
anne.baillot@univ-lemans.fr, francois.vignale@univ-lemans.fr

What is READ-IT?

This paper aims to present READ-IT (Reading Europe Advanced Data Investigation Tool) and argue, based on this example, that the management of big data produced by Cultural Heritage Research has the potential to be at the core of the development of multimodal research in the Humanities in the years to come. READ-IT is a 3-years transnational and interdisciplinary research project awarded in December 2017 by the Joint Programming Initiative for Cultural Heritage[1]. It started officially on June 1st, 2018 and will end in 2021.

The goal of the READ-IT project is to build a unique large-scale, user-friendly, open access, semantically-enriched investigation tool to identify and share groundbreaking evidence about 18th-21st century Cultural Heritage of reading in Europe. READ-IT will investigate innovative ways of gathering new resources through crowd-sourcing and web-crawling as well as linking and reusing preexisting datasets, and try to define a common denominator of European reading experiences across times and cultures. READ-IT will thus ensure the sustainable and reusable aggregation of qualitative data allowing an in-depth analysis of the Cultural Heritage of reading.

In terms of its structure, READ-IT consists in a consortium of 5 academic partners from 4 different countries (Institute of Czech Literature, Czech Academy of Sciences; Open University, London; Utrecht University, Netherlands; CNRS-IRISA, Rennes, France and Le Mans University for France). The consortium is led by Le Mans University (Professor Brigitte Ouvry-Vial). The Institute of Czech Literature coordinates the use cases and Open University the dissemination actions, thus taking a significant part in the definition of the ontology which is at the core of the project. Utrecht University is in charge of the web interface, CNRS-IRISA is taking care of the core computer science research and Le Mans Université is managing the project, cataloging requirements and taking part in the definition of the ontology, the data model and the vocabularies associated. The complete research team consists in a pool of 16 scholars[2].

1 http://www.jpi-culturalheritage.eu/

2 READ-IT has been inspired by the 2006 pioneering manually entered UK-Reading Experience Database (http://www.open.ac.uk/Arts/reading/UK) and by the EuRED proof-of-concept database

READ-IT and history of reading practices.

Regarding history of reading practices, knowledge has significantly increased over the last decades about what, where and when people read. Still, two major questions remain unanswered: why and how do people read?

These two main research questions can also be decomposed in a series of subordinated questions:

- What kind of transaction does there exist between a reader and a text?
- What role does the environment play in this transaction?
- Is it possible to list and model the emotions caused by reading?
- Have these emotions changed throughout time and space in Europe?
- Is it possible to sketch out the portrait of something like the « European reader »?

In order to provide an answer to these questions, one must consider the act of reading as an individual experience with modalities specific of each iteration of the experience, and not following one single pattern. This approach also implies that one must take into account not only the reading testimony itself, but also the circumstances or the context that accompany its production.

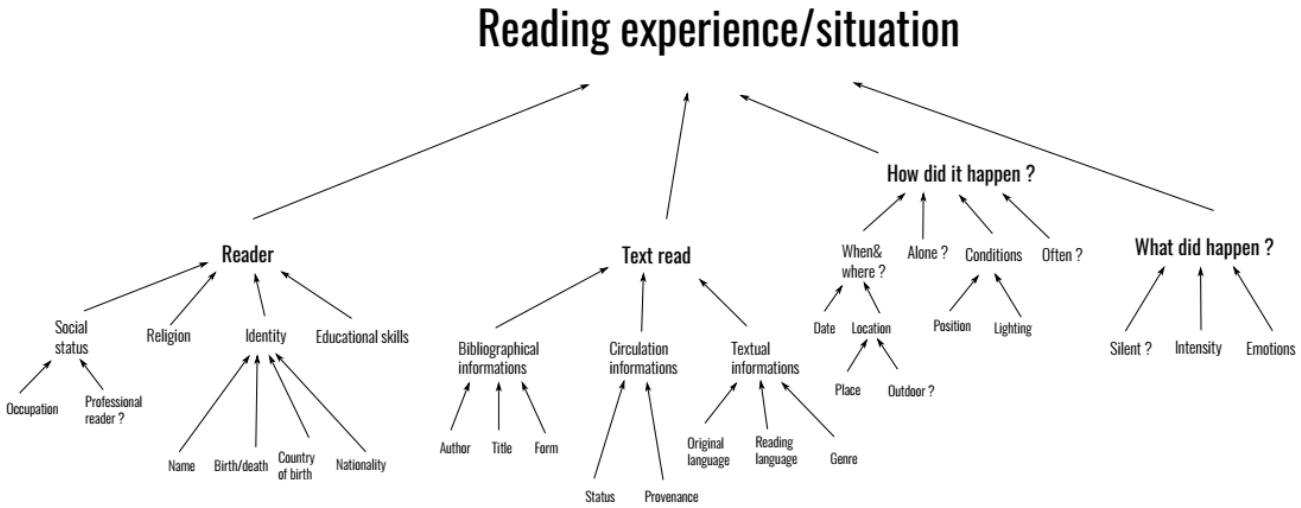
However, in doing so, one faces a major methodological issue, namely that reading is a brain activity that leaves no direct or tangible traces. Within the READ-IT project, the research team thus has to deal with direct or indirect testimonies, concomitant of the reading experience or memories which, in addition to their great heterogeneity in terms of reliability and completeness, will take the form of representations, narratives or reports. The major challenge here is to find a structure that allows to grasp different narratives, different modalities, different temporalities, and still be able to aggregate them in a fruitful way in order to address the research questions listed above.

Defining reading experiences and reading situations

Within READ-IT, we make a difference between *reading experience* and *reading situation*. This distinction is based on the media through which the experience or situation is transcribed or depicted. We speak of an *experience* when the testimony consists in a narrative or a report. This occurs mainly in textual sources and audio or video recordings. Conversely, we speak of a *situation* when the trace consists in a representation such as an image, painting, photograph, engraving …

Broadly speaking, at this preliminary stage, we consider that a reading experience (or situation) is composed of 4 large groups of items. An experience (or situation) engages 1) a reader, 2) a text read in a certain context that can give rise to 3) certain modalities and 4) certain consequences or results. Each of these items can then be subdivided into lower levels according to the degree of granularity required to avoid ambiguities and ensure data quality. Information sought about the reader are mainly biographical (name, date of birth, place of birth, social status, religion, …) while information sought about the text read are mostly bibliographical (title, author, form, language, provenance, …). Circumstantial or contextual elements will inform about the date and place of the experience, whether it was public or solitary, about the material conditions (lighting, position), whether it was repeated or not.

Elements related to the effects or consequences of the reading will inform on its silent (or not silent) character, on its intensity and especially on the emotions caused by the reading.



Reader and text read answer the “what” question, while circumstances and effects address the “how. One of the major elements of the characterization of any reading experience or situation is that we always have in the one hand a reader and a text (which corresponds to the “what) and on the other hand circumstances (the “how).

This combination leads to a theoretical – and minimal - definition of the reading experience/situation as a testimony (or trace) that provides at least one piece of information regarding the what **and** one piece of information regarding the how. In practical terms, the combination or association of at least two of the items listed above - as long as they belong to the what and the how- is potentially a reading experience/situation and, therefore can be indexed in the database through a data model and an ontology.

Workplan and sources

The workplan in READ-IT follows two purposes: 1) identifying and describing reading experiences/situations; 2) connecting reading experiences/situations with their context.

The sources used or reused within READ-IT are multimodal and consist of textual sources (digitized but not necessarily ocr-ized), images (paintings, photographs, drawings, engravings …), audio recordings (transcribed), digital-born content (blogs, social medias). These sources have to be in the public domain or CC-like, already digitized, FAIR data principles compliant, available in English, French, German, Dutch, Italian or Czech and they must be dated from the 18th to the 21st century.

The two examples below show the kind of information READ-IT wants to retrieve, extract and index.

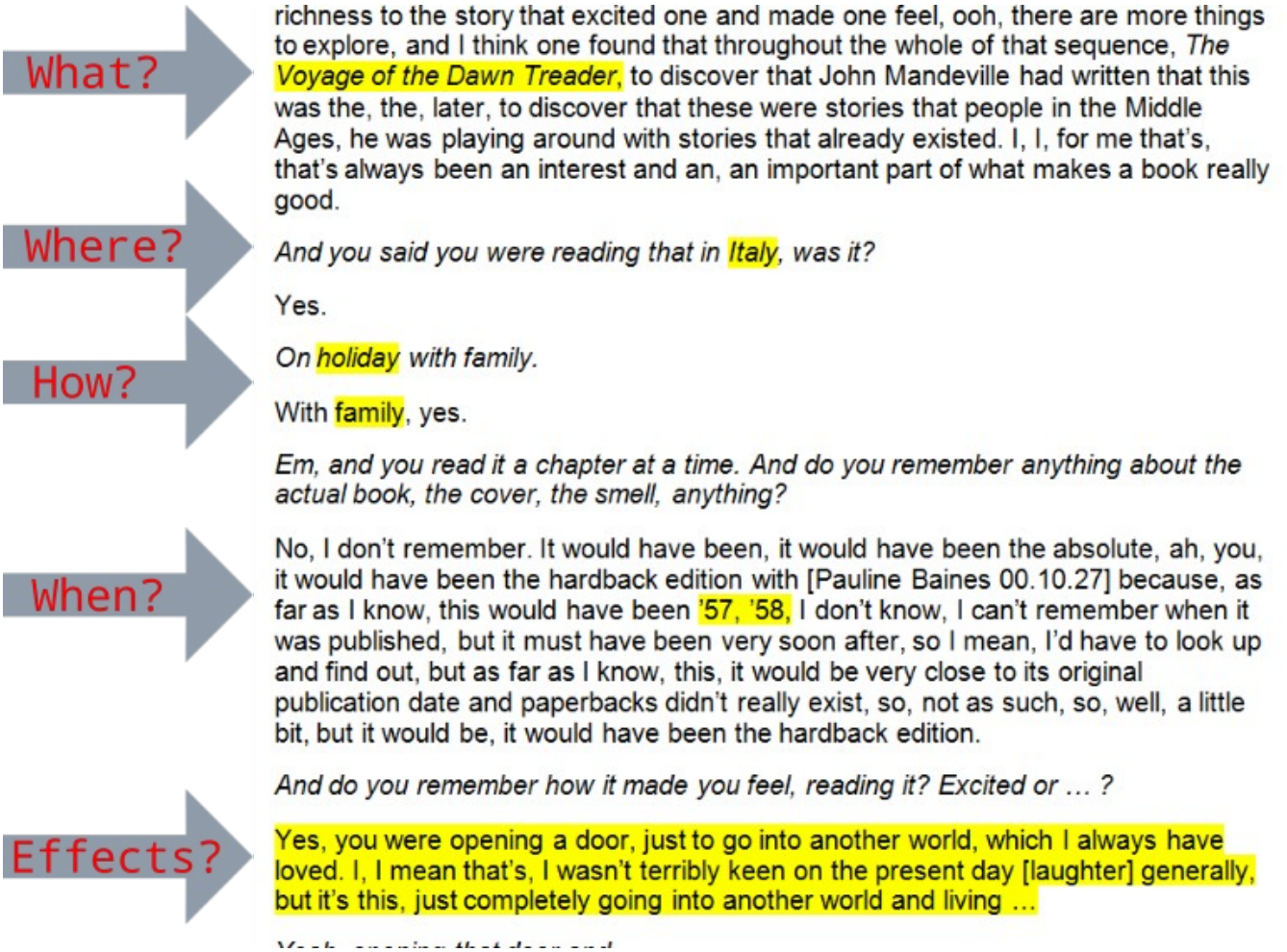


Illustration 1: Memories of fiction, Ferelith H. interview part 2 [transcribed], 2014, https://soundcloud.com/memoriesoffiction/ferelith-part-2

The extract presented here is part of a transcribed interview that can be found in an oral history project called "Memories of fiction" where people are asked about their reading practices. A corpus like this one is of particular interest to the READ-IT team in order to address the challenge of automatically identifying scattered pieces of information that, once they will be gathered together, could lead to the characterization of a reading experience. In this example, we have a text read (what?): *Voyage of the Dawn treader*; a location where the experience took place (where?): Italy; circumstances (how?): on holiday with family; a date (when?): not before 1957 and some feelings expressed by the reader (effects?): "go to another world, which I have always loved".

After this identification phase, the next major technological challenge of READ-IT is to ensure that these dispersed elements, when confronted to the ontology and the data model, can be gathered and indexed in a unique consistent record.

An example illustrating this work step can be elaborated on based on the painting by Mary Cassatt below.

Illustration 2: Mary Cassatt. Nurse reading to a little girl, 1895, Metropolitan Museum of Art, New-York

When it comes to images, the method is slightly different. In this case, the algorithms developed within the project will be required to recognize some pieces of information that could ultimately lead to the characterization of a reading situation, but due to the filter of the artist's perception, only a certain degree of certainty can be achieved. Here, the major challenge consists in making the computer recognize a potential interaction between one or several human beings and a written support, and only then to propose a maximum of description elements. In this painting, we can see two individuals, seating outside in the daylight, possibly holding a book - and for that reason, maybe reading.

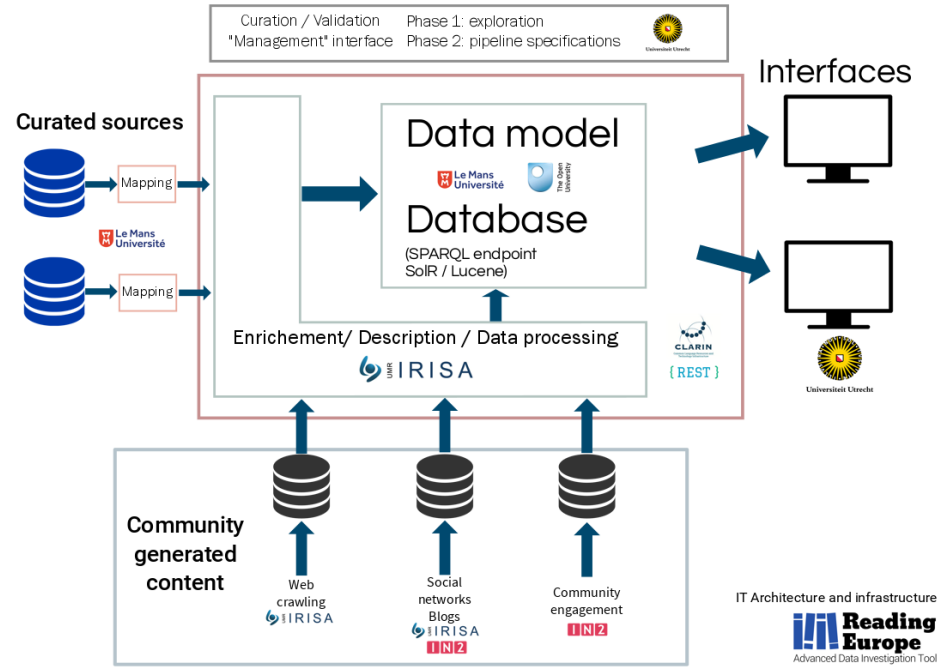
As a consequence, computers are only asked to identify candidate images. After that, human beings brought together via crowd-sourcing campaigns will validate or not the machine's interpretation and then contribute to its training.

Technological foundations

Technically speaking, the IT architecture and infrastructure follows the schema above. Two different kinds of sources are envisioned: community generated content and, more important here, curated content emanating from cultural institutions or other research projects (Memories of Fiction, UK-RED …).

The goal is to gather and connect state-of-the-art technology in semantic web, linked data and multimedia analytics in order to provide structured descriptions of reading experience resources and to facilitate reuse and interoperability with related web data across languages. The choice of technologies and standards, such as RDF and SPARQL, combined with ontology formalisms and widespread vocabularies, allow READ-IT to publish its structured data based on a highly expressive and universally understood data model. This will facilitate the conception and implementation of flexible yet powerful interfaces for search and exploration

purposes, allowing groups of users with different research agendas to engage with READ-IT resources. Text mining and multimedia information retrieval are intended to be extensively exploited: natural language processing (NLP) will play a pivotal role in identifying new sources of information on the web; multimodal content analysis and image retrieval will enable the curation of rich multimodal resources; entity detection and linking in key languages (English, French, German, Italian) will facilitate the description and curation of data.



READ-IT and cultural heritage institutions

The realization of this ambitious endeavor depends on the availability of curated digital collections. The three major reasons for this desideratum are the following. First, curated collections are organized, described and often indexed. In addition to that, Intellectual Property Rights issues are most often dealt with. Second, these collections are more than likely to contain the traces or the testimonies needed in READ-IT in order to construct part of the answer to the research questions we raised about the how and the what that we want to answer. Third, READ-IT needs a considerable amount of multimodal data to train the machine and improve the performance of the algorithms that will be implemented.

In exchange to the access to this digital heritage of reading, this project aims at helping cultural institutions such as libraries or museums to highlight their own collections by giving its users the possibility to annotate and enrich their collections (making use of the possibility to create special digital collections about reading within the READ-IT portal, or building one's own). APIs developed within READ-IT will be made open source-available. They should be made replicable and transposable to other esthetical experiences such as music or artworks.

Acknowledgments.

This work has been funded by the Agence Nationale de la Recherche (ANR-17-JPCH-0001-01).