Understanding the phenomenology of reading through modelling

Alessio Antonini, Mari Carmen Suárez-Figueroa, Alessandro Adamou, Francesca Benatti, François Vignale, Guillaume Gravier, Lucia Lupi, Brigitte Ouvry-Vial

Understanding the phenomenology of reading through modelling

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Alessio Antonini[a,*], Mari Carmen Suárez-Figueroa[b], Alessandro Adamou[c], Francesca Benatti[d], François Vignale[e], Guillaume Gravier[f] and Lucia Lupi[g]
[a]*Knowledge Media Institute (KMi), The Open University (OU), MK7 6AA, Milton Keynes, UK*
[b]*Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM), Campus de Montegancedo sn, 28660, Madrid, ES*
[c]*Insight Centre for Data Analytics, NUI Galway (NUIG), IDA business park, Lower Dangan, Galway, IE*
[d]*Department of English and Creative Writing, The Open University (OU), MK7 6AA, Milton Keynes, UK*
[e]*Le Mans Université, Avenue Olivier Messiaen, 72085 Le Mans, FR*
[f]*Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), 263 Avenue Général Leclerc, 35000 Rennes, FR*
[g]*Dipartimento Interateneo di Scienze, Progetto e Politiche del Territorio (DIST), Polytechnic of Turin & Univerisity of Turin, Viale Mattioli, 39, 10125 Torino, IT*

Abstract. Large scale cultural heritage datasets and computational methods for the humanities research framework are the two pillars of Digital Humanities, a research field aiming to expand humanities studies beyond specific sources and periods to address macroscope research questions on broad human phenomena. In this regard, the development of machine-readable semantically enriched data models based on a cross-disciplinary "language" of phenomena is critical for achieving the interoperability of research data. This contribution reports, documents, and discusses the development of a model for the study of reading experiences as part of the EU JPI-CH project Reading Europe Advanced Data Investigation Tool (READ-IT). Through the discussion of the READ-IT ontology of reading experience, this contribution will highlight and address three challenges emerging from the development of a conceptual model for the support of research on cultural heritage. Firstly, this contribution addresses modelling for multi-disciplinary research. Secondly, this work addresses the development of an ontology of reading experience, under the light of the experience of previous projects, and of ongoing and future research developments. Lastly, this contribution addresses the validation of a conceptual model in the context of ongoing research, the lack of a consolidated set of theories and of a consensus of domain experts.

Keywords: Reading Experience, Conceptual Modelling, Experience Ontology, Digital Humanities, Modelling Methods

1. Introduction

The combination of digital sources and computational methods is at the centre of a change of paradigm and of research breakthroughs on cultural heritage. Firstly, the discoverability of sources described and enriched through the Semantic Web enables the construction of integrated datasets of sources based on different archives. Secondly, data integration, quantitative and qualitative studies complement the in-depth analysis of individual sources. The use of large-scale datasets and computational methods applied within a humanities research framework is the pillar of the revolution of the Digital Humanities.

* Corresponding author. E-mail: alessio.antonini@open.ac.uk.

The current challenge for the Digital Humanities is how to scale up from the established paradigm of focused studies of specific sources and periods, to macroscope research addressing broad human phenomena over the *longue durée* as represented in the cultural heritage [1]. While the humanistic research of the past has focused on scarce and hence exceptional case studies, the radical digital reconstruction of the cultural heritage archive permits for the first time the study of more extensive contexts or ideas [2]. In this vision, the interplay between the systematic study of data generated by research case studies sharing a general focus on a human phenomenon could unlock advancements on macro-scale questions related to understanding the human condition through time [3]. To realise this vision, a crucial issue to be addressed is the interoperability of research data and therefore the development of a shared "language" for the formalisation of a given phenomenon to be used in the production of computable research data [4]. Lastly, to enable the use of computational methods of the Digital Humanities, research data must be machine-readable and include semantically enriched extensions of the descriptions of cultural heritage artefacts, possibly by automatic means, so that contextual knowledge about them can be retained [5].

This contribution reports, documents and discusses the development of a model for the study of reading experiences. The modelling of the reading experience is part of the EU JPI-CH project Reading Europe Advanced Data Investigation Tool (READ-IT)[1]. The approach of the READ-IT project is based on the development of a technological ecosystem, including data collection tools and annotation algorithms, populating a common database about reading experiences, via the development of multidisciplinary research case studies. With the help of the study of data collected in the READ-IT database, the READ-IT consortium aims to address macroscope questions regarding the evolution of reading in Europe during the last three centuries.

By referring to the experience of READ-IT, this article will provide a discussion of the challenges related to the development of a semantic model in the framework of the Digital Humanities with the focus on a complex phenomenon such as reading. Specifically, this contribution will address:

- the approach to conceptual modelling in a multidisciplinary framework
- the modelling of the reading experience phenomenon in the light of previous projects and existing standards
- the strategy for the validation of the conceptual model aimed at supporting the generation of new data and the development of new tools for supporting research activities.

Through the discussion of READ-IT, the authors will highlight three orders of challenges emerging from the development of a conceptual model for the support of research on cultural heritage. The first order of challenges emerges from the limits of existing research data, grounded on the specific research framework of individual case studies. The second order of challenges is related to the presence of a vast landscape of disciplinary theories addressing specific aspects of the reading phenomenon. The third order of challenges is related to the requirements emerging from the research activities. Based on the discussion of the emerging challenges, the contribution documents the rationale behind the methodology developed, the specific modelling choices, and the validation of the model of reading experience.

The rest of the paper is structured as follows:

- Section 2 - 2. Ontology Development Approach, description of the modelling process
- Section 3 - Previous Projects, discussion of related previous projects
- Section 4 – Ontology Requirements, description of ontology requirements
- Section 5 - Reused Ontologies, brief introduction of the reused ontologies
- Section 6 - Reading Experience Ontology, background and description of the ontology
- Section 7 - Validation, discussion about how the ontology addresses the requirements and supports the research on reading
- Section 8 - State of the Art & Related Work, state of the art on ontology validation and brief presentation of relevant ontologies
- Section 9 - Conclusions, brief description of the contribution and open issues, outline of the next developments of the ontology.

2. Ontology Development Approach

The theoretical frame of the READ-IT project did not include a consensus or a theory of reading that integrates the perspectives of the different research groups. Furthermore, the conceptual modelling of the reading experience had to take other constraints related to legacy data of previous projects, the data

[1] https://readit-project.eu/

collection for a wide range of disciplinary and interdisciplinary case studies and the concurrent development of new tools.

In this scenario, we developed a specific modelling methodology motivated by sources, informed by theory, validated through case studies, iterative, incremental and engaged with the different project partners. Specifically, the modelling lifecycle was structured in four main phases, see Fig. 1.

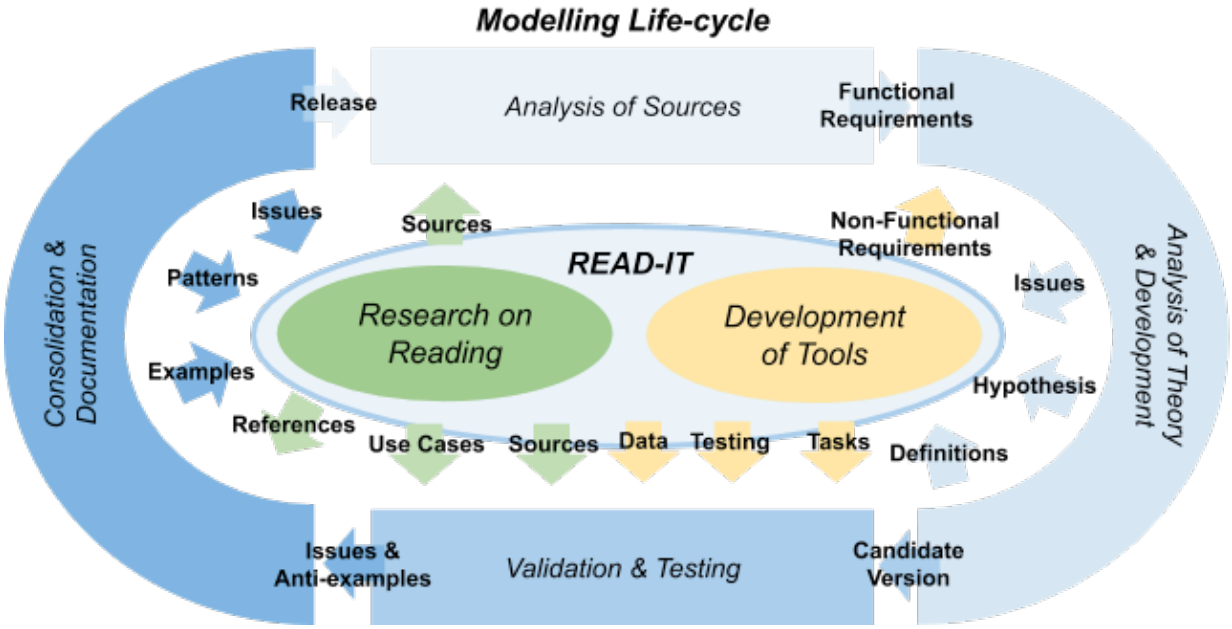

Fig. 1. The modelling lifecycle main phases and interactions with research and development work strands.

2.1. Analysis of Sources

The analysis of the sources used to ground the concepts of the model is based on the experiences of reading documented in the sources. The sources are provided by researchers involved in READ-IT and represent a significative set of the different types of sources and reading experiences. This phase is used to define concepts and functional requirements to fed into the development phase.

2.2. Analysis of Theory & Development of the Model

The study of theories about reading, experience and action is used to guide the integration of concepts and to fill the gaps in the examples included in the sources. The development of the model, informed by the theory, addresses the formalisation of concepts and structures [6]. The development takes into account the functional requirements defined during the analysis of sources and the non-functional requirements emerging from the design and development of the tools. The development phase aims at producing a candidate model for the next phase, and its objectives are the identification of issues and hypotheses that will be assessed with the help of researchers and consolidate the state of the working definitions (of key concepts) of the project.

2.3. Validation & Testing

The validity of the model is defined as the ability to encode the relevant facets of the reading experience in the sources and to provide and support the data-related research (see Section 7 - Validation). In this regard, the validation is performed through the engagement of researchers on reading and development teams. The engagement of researchers addresses the details of their case studies, the use cases and the annotation of new sources. The engagement of technical partners involved the discussion of the tasks related to the tools which are relevant for the model, the testing of tools and the discussion of data-related issues. The outputs of this phase are new tasks for the backlog related to anti-examples (i.e. documented unwanted ontological commitments) and examples

that are yet to be addressed (e.g. related to the test sources).

2.4. Consolidation & Documentation

The consolidation phase will address the open issues while generating documentation including examples, design patterns and highlighting the issues that can be addressed only as hypothesis or that require a contribution from the research strand of work.

3. Previous Projects

The design of the READ-IT ontology is based on the experience and limitations of previous projects in cataloguing experience recorded in literature.

3.1. UK-RED

The UK-RED (UK Reading Experience Database)[2] ontology was developed to support Digital Humanities research on reading experience. The RED ontology and database were the result of an incremental rationalisation of research data produced in multiple projects and through the engagement of Humanities students and volunteers.

The RED ontology [3] includes three classes: foaf:Person, foaf:Document and red:Experience. RED used concepts from the linked event (event), Friend-of-a-Friend[4] (foaf) and DBpedia ontologies[5] (dbo). The pillar of RED is the concept of reading Experience, connecting a foaf:Person in the role of reader or listener, place and date of the reading, and a document (object of reading).

The RED ontology is supported by 10 years of use. A pioneering project, RED was based on sources collected directly by researchers and stored in RED. Thus, RED does not have connections with external repositories. Furthermore, the conceptualisation of Experience was scoped on the few objective facts related to experience: reader (agent), document (object of reading), time and location of reading. The simplicity of the RED model of reading allowed the accumulation in RED of a heterogeneous collection of sources and research data. On the other hand, the simple schema encoding the research data limits the reuse to RED content as a repository of sources about reading experience, rather than a repository of research data about reading experience.

3.2. LED

Similar to UK-RED, the aim of the LED (Listening Experience Database)[6] ontology is to support research on listening experience. The LED strategy is two-fold: 1) reconstructing the context of the listening experience (linking places, performers, events, musical works, etc.), and 2) supporting the collaborative distributed incremental curation of sources about listening experience.

The LED ontologies (led) revolve around the concept of (listening) experience as "a documented engagement of an individual in an event of one or more pieces of music being performed" [7]. In this regard, LED core can be regarded as connecting four main notions: Listening Experience, Source, Agent and Music. LED considers the listening experiences themselves to be events, if subjectively perceived ones, therefore relies upon the existing event ontology literature: as a result, it treats the participants to an experience as agents, its source document as a report and the music being heard as event factors. Sources are for the most part described from a bibliographical take using the BIBO ontology[7], integrating it with a controlled vocabulary of source categories typically not formalised in structured datasets (such as oral history or official papers). As musical performances and works are represented using the FRBR-aligned Music Ontology[8], so too are the roles of participants in an experience modelled after that ontology, combined with BIBO. This makes it possible to query the datasets to find e.g. authors who reported on experiences where they played the music themselves.

As an evolution of RED, LED structures and includes a formal description of the curation process of sources as part of the database and approaches the representation of the experiences as an abstraction of external repositories (result of the curation process), not limited to first-hand collected sources. Although LED does provide an enriched description of the facts related to the experience, it still does not describe their phenomenology.

[2] http://www.open.ac.uk/Arts/reading/UK/
[3] http://data.open.ac.uk/page/context/red
[4] http://xmlns.com/foaf/spec/
[5] http://dbpedia.org/ontology/

[6] https://led.kmi.open.ac.uk/linkeddata/
[7] http://bibliontology.com
[8] http://musicontology.com/specification/

4. Ontology Requirements

READ-IT builds on the experience of previous Digital Humanities research on cultural experiences[9], with the aim of overcoming a number of crucial conceptual and technical limitations of these projects. These include for example their limited geographical range and their lack of integration of multimodal digitized sources of evidence (images, texts, audio sources, computer-mediated communications). The READ-IT information architecture, see Fig. 2, is an evolution of previous projects, taking into account their limitations and the new requirements related to the envisioned interoperability of research data and scaling up of Digital Humanities research.

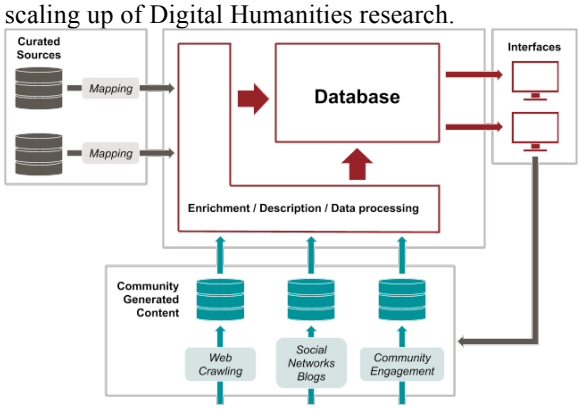

Fig. 2. READ-IT general architecture. The model of reading guides the design of the database, the systematic annotation of the curated sources and the analysis of the community-generated content, and it is an important factor in the design of the user interface.

In the READ-IT framework, the model of reading has the role of informing the design of 1) the database that should integrate research data produced through case studies and 2) the tools that will support research activities (e.g. annotation algorithms, annotation tools, crowdsourcing tools). Last but not least, the model of reading will be the main resource to guide the reuse of research data, thus providing a common language about the phenomenon of reading and guiding the use of the database, see Fig. 3.

4.1. Non-functional Requirements

This scenario indicates a set of requirements related to the direct role of the model in the READ-IT architecture and its indirect role in the READ-IT research activities.

[9] Such as the UK-Reading Experience Database (2006-present), the Reading in Europe: Contemporary Issues in Historical and Comparative Perspectives project (2014-2017) and the Listening Experience Database project (2012-present)

4.1.1. Types of Sources

The primary source of reading experiences are cultural heritage sources. However, new research strands on digital reading are focusing on hypertext, social media, e-readers, web-novels, webcomics, podcasts, and new emerging media. In this scenario, we have to consider a wide and open range of source types as sources providing insight on reading experience.

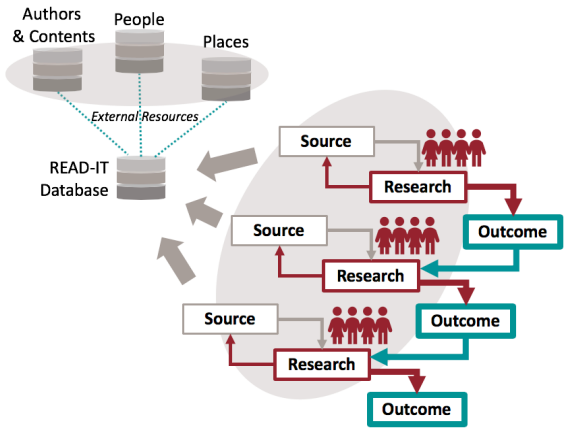

Fig. 3. READ-IT Database is going to collect sources and annotations produced through case studies (outcomes) their reuse in the development of new case studies.

4.1.2. Research Data

The potential focus of research on reading is a wide landscape, ranging from classical reception, narrative reception, expert reading, interaction with text, cognitive effects, history of reading to digital reading. The research initiatives addressing reading follow a wide range of methodologies producing data about different aspects of the reading experience. Regardless of the specifics of research activities, we consider exclusively annotations of sources and sources (Research Data) about reading.

4.1.3. Use of the Ontology

As the different research initiatives, the research data about different aspects of reading experiences may represent a specific facet of the reading experience. In this regard, we must consider a partial use of the ontology in the annotation of sources, as well as allowing flexibility in the model design.

4.1.4. Data Integration

Research data produced by different case studies and research initiatives will focus on different facets of the reading phenomenon (partial data). Research dataset only partially representing the phenomenon

of reading could be difficult to reuse outside the same research framework. Thus, the ontology should support the reconstruction of the phenomenon of reading as a whole and the integration of partial datasets about specific aspects of reading.

4.1.5. Training set

Research data will be used to train machine learning algorithms for the automatic annotation of sources. To be used as training sets, the research data must be integrated by making explicit derivable facts and incorporating validation of annotations by experts to differentiate them from automatic annotations.

4.2. Functional Requirements

Though the support of the researchers involved in READ-IT, we collected a sample of sources about reading experiences. The sampling of sources had been selected to be representative of the different types of sources, e.g. transcripts of interviews, diaries and to highlight the richness and complexity in terms of the description of the reading experience, e.g. comparative reading, multiple readers, re-reading.

The functional requirements had been identified primarily through the analysis of sources and secondarily through the direct engagement of the humanities researchers involved in READ-IT. The analysis of sources aimed to identify the core concepts emerging from the descriptions of the experience of reading. The engagement of researchers on reading had the aim of discovering gaps in the identified concepts. The analysis was conducted in two cycles, during the first six months of the project, during which the following clusters of requirements were identified.

4.2.1. Reading activity

Readers identify an aspect of the activity of reading as causing a specific effect they experienced, e.g. an emotion, a memory, a judgement. The analysis of the different cases showed that the effects of reading are related to: 1) a moment, e.g. related to a twist in the story or setting, related to a specific segment of the content 2) an episode of reading, e.g. on a train, a bedtime story, related to its contingency, and 3) a whole reading, e.g. reading of *War and Peace*, related to the interaction between reader and content. Summarising, the articulation of the reading activity is of relevance for the description of cause / effect relations between reading and readers' conditions.

4.2.2. Readers' Conditions

The description of the readers' conditions includes information about 1) the human/social situation of reading and 2) their personal mental state. In the description of the situation, the focus is mostly at the social scale, e.g. "on the train coming back from my grandmother, outside Oxford". The details about place, time and material conditions of reading are characteristics of a socially relevant event, e.g. "visit to grandmother", "vacation with family". Secondly, readers describe specific aspects of their mental state triggered by reading. The readers' descriptions focus on the changes of a single specific aspect of their mental conditions, e.g. "I felt excited", "I was swallowed by the reading". Summarising, readers' conditions are related to the social context of reading, in which the reader could have either a passive role (e.g. climatic event) or an active role (e.g. vacation), and to the specific aspects of their mental state changing in reaction to reading.

4.2.3. Effects of Reading

The analysis of the collection of reading experiences outlines a landscape of facts, ideas, concepts, judgements that readers use, refer to and consider. In the readers' narrative, reading (and its cumulative effects) takes the role of baseline for the description of the effects of other readers, e.g. "unlike in the first book, in the second book the authors fail to address the condition of women". This landscape is the result of incremental and cumulative effects of reading. Readers report about the evolution of their evaluation of contents, in relation to re-reading or time to think and re-think about it. Summarising, the effects of reading can be: a) a direct consequences of a specific aspect of the reading activity, b) a cumulative result of multiple readings or c) an incremental result of further reflection outside the scope of the reading activity.

4.2.4. Approach to Reading

Readers describe their standpoints in approaching reading from two main perspectives. Firstly, readers ground their stance on their experience and personal or social condition, e.g. previous judgements about topics, authors, contents. Secondly, readers refer to their stance in terms of developed reading habits, e.g. how often, where, when, why, what they read. In the first case, readers take a personal perspective about what they like or dislike, what they wish to accomplish or expect and the activities in which they are involved. In the second case, readers take an external

perspective on their behavioural patterns, e.g. "this is always present in the books I like", "I read it every day to my kids". Summarising, the description of the readers' approach to reading is an important component in the description of the experience. Readers can provide a perspective on the frame of mind in which they approach reading, or the background of their typical readings. In both cases, the approach to reading is entangled with contingent condition of the reader at the time of reading, e.g. "I was sixty-six", "I was in second grade".

5. Reused Ontologies

Ontology reuse has traditionally been considered a basic activity and a best practice in Ontology Engineering. In the context of the NeOn Methodology [8,9] prescriptive methodological guidelines for reusing ontologies have been proposed. These guidelines cover the following activities: (1) search repositories for candidate ontologies that could satisfy the needs of the ontology to be developed; (2) assess whether the set of candidate ontologies are useful for building the ontology; (3) select the best candidate ontologies for developing the ontology on the basis of a set of criteria; and, (4) integrate the selected ontologies into the one being built.

Following this approach and based on the analysis of the ontology requirements, we assessed existing related ontologies. The criteria used were concern to 1) management of annotations and sources and 2) the description of the content of sources, with special focus on human agents and creative contents.

5.1. W3C Web Annotation Data Model

The W3C Web Annotation Data Model [10] (WADM) is a schema describing the concepts and relations between an annotation (e.g. comment) and a web source. W3C WADM is used to represent the annotation of cultural heritage sources and other sources used in the research use cases, enabling the highlighting of specific content excerpts as the description of specific aspects of the reading experience. Specifically, the W3C WADM concepts annotation, body and target are used to encode the agent responsible and software used for creating the annotation, the graph representing the reading experience described in the source, and the selection of the source target of the annotation.

5.2. CIDOC CRM Ecosystem

"The CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation" [10]. CIDOC-CRM is an official standard, ISO 21127:2006, and therefore the reference vocabulary for digital representation of cultural heritage. CIDOC CRM is a core ontology for a family of specialist ontologies, such as FRBRoo[11] and CRMsoc[12].

5.2.1. CIDOC CRM

The CIDOC CRM core provides the conceptual backbone of the READ-IT ontology of reading experience. Indeed, CIDOC CRM concepts of temporal entity, event, condition state and activity are at the core of the ontology. For instance, the structure of the reading process is built exploiting the features of activity, event and temporal entity to describe the process of reading articulated in multiple sessions of different duration occurring in different places, involving the participation of multiple people and interconnected in a progression.

5.2.2. FRBRoo

The Functional Requirements for Bibliographic Records model (FRBR) is a conceptual model of bibliographic contents. FRBR provides a conceptualisation of the life-cycle of creative contents: author's *work* (e.g. a romance), the different forms of *expression* of work (e.g. a version of the romance), the different forms *manifestation* of a work expression (e.g. material prototype of a book) and lastly the several *items* which are instances of a specific manifestation (e.g. book copies).

FRBRoo is an ontology of the CIDOC CRM ecosystem encoding the concepts of FRBR.

5.2.3. CRMsoc

"CRM Social is a domain ontology extending CIDOC CRM aimed to support and capture social documentation" [11]. CRMsoc has not been released yet but, nevertheless, we take it into account.

CRMsoc (socE) is expected to provide a standard solution to the representation of concepts such as social status, political stance, gender, which are of

[10] https://www.w3.org/TR/annotation-model/

[11] http://www.cidoc-crm.org/frbroo/home-0

[12] http://www.cidoc-crm.org/crmsoc/ModelVersion/version-0.1

great relevance for the description of the reader but outside the scope of the research on reading.

On the other hand, the socE Mental Attitude class overlaps with the READ-IT class State of Mind. The socE Mental Attitude represents the intention related to a plan and reading is indeed a specific type of an activity, implementation of a plan, supported by reader's intentionality.

At the current stage of socE, socE Mental Attitude as its definition is still an open issue. In the current documentation, socE Mental Attitude is described as "conscious maintaining of an intellectual attitude towards matters of knowing, believing or guiding actions and reactions to social and other environmental situations, such as, besides others, beliefs about laws governing nature or intentions to carry out actions" [12]. Therefore, in the READ-IT domain, socE Mental Attitude is close to the role of State of Mind as premise of reading and to the reader's disposition toward reading, medium or content.

The class socE Intention to Apply, specialisation of scoE Mental Attitude, is related to the concepts of E39 Actor and socE Activity Plan as the intention of an actor toward of implementing a plan. In this frame, socE Intention to Apply could be used to encode the intention of reading, as a specialisation of State of Mind preceding and premise of a reading process.

5.3. FOAF

Friend of a friend (FOAF) is an ontology addressing the digital representation of people and the relations between people and web contents. FOAF is used to describe the reader and participants involved directly or indirectly in the situation of reading.

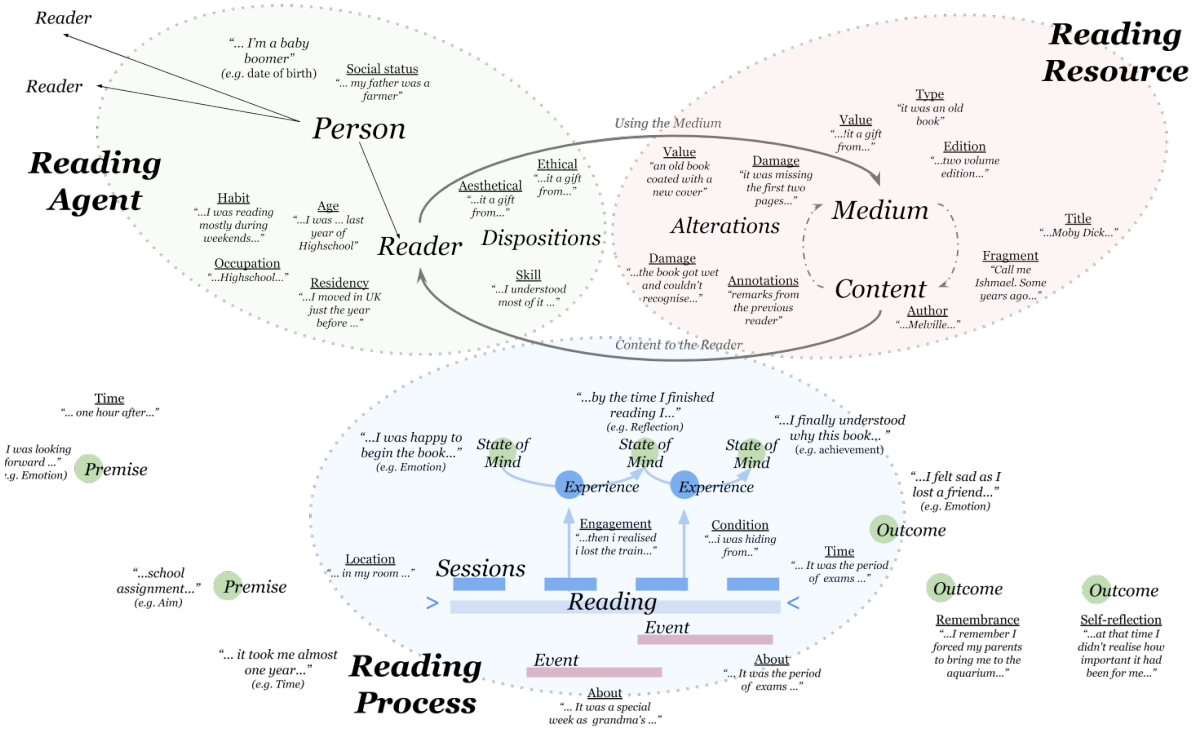

Fig. 4. The three main conceptual areas addressed by the READ-IT ontology of reading experience, main concepts, relations and properties.

6. Reading Experience Ontology

The design of the READ-IT ontology focuses on addressing the description of the human experience of reading. The aim of the

ontology is to provide the common language to structure and share research data addressing aspects related to three main conceptual areas (see

Fig. 4):

1. the *reading agent*, who is reading, why and what are their conditions approaching a reading;
2. the *reading resource*, what is being read, what is the type and condition of the medium of reading;
3. the *process of reading*, where and when reading takes place, in which material and social conditions and how a reading is carried out.

The ontology will address the interactions between these three systems: the situated interactions between reader, medium and content, and the effects of reading during and after reading.

In the following section, the concepts of the READ-IT ontology are described using local names, e.g. Reader, Medium, and as prefixed q-names, e.g. foaf:Person, and with terms from other third-party ontologies, specifically: Friend of a Friend (foaf), CIDOC CRM (cdc) and FRBRoo (frbr). The diagrams follow the RDF graph notation.

6.1. Source and Annotation

The reading experiences we are representing with the experience ontology are annotations made on a variety of sources. The reason why we rely on annotation rather than on categorization is to enable researchers to address a wide range of problems in reading studies. In READ-IT we adopt the W3C web annotation data model (wadm) to represent the annotation process. This model to connects a graph (wadm:body) to the metadata describing the annotation process (wadm:annotation), e.g. annotation agent, with the source of the annotation, wadm:target.

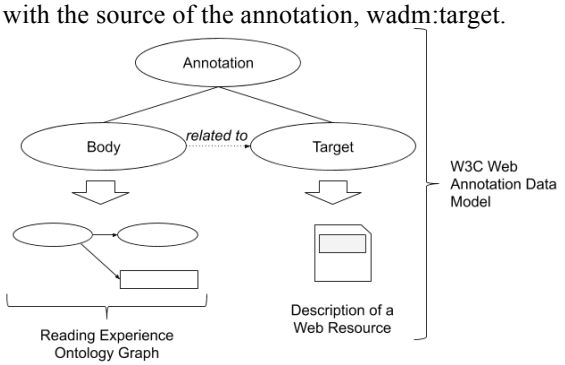

Fig. 5. Relationship between W3C Web Annotation Data Model and reading experience ontology graphs.

The findings about a reading experience in a source, i.e. the body of an annotation, is a sub-graph generated using the reading experience ontology, see Fig. 5. Thus, a research dataset is a collection of wadm graphs, annotation, body and target and relative graphs about reading experience.

The rest of the article addresses the value of the wadm:body and how to structure the reading experience, while it omits information about the source and the annotation.

6.2. Approach to Reading

The analysis of sources highlighted a set of requirements and concepts related to reading experience. As argued, concepts and requirements report about different aspects of reading: a fragmented view of the phenomenon. In this scenario, the modelling of the reading experience required the integration of the different aspects of reading. These gaps have been addressed through the analysis of existing theories about reading, action and experience.

The analysis of theories of reading indicated the existence of an underlying dynamic connecting the different aspects of reading. Specifically, the modelling of the reading experience was grounded on the following assumptions: 1) a reading process triggers a process of sense-making defining, as a result, the reading in the reader's mind (from an ongoing process to a concluded event); 2) reading is grounded on the experience of the reader, in terms of previous reading and, in general, events, see Fig. 6.

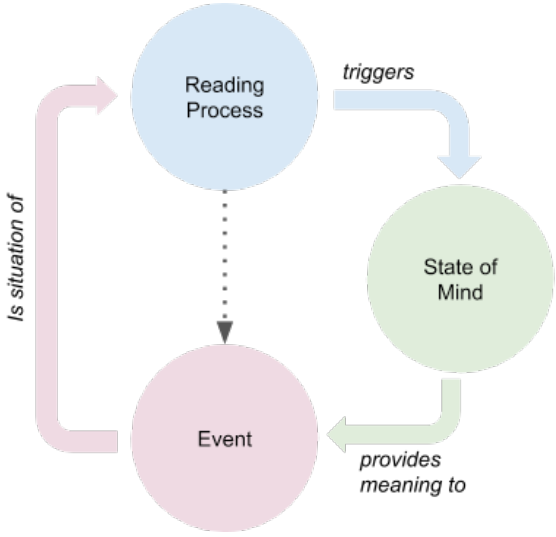

Fig. 6. Connections between activity, experience and event.

6.3. Agent, Resource and Process

The analysis of sources guided the identification of the core concepts of the ontology. Similarly to LED and RED, the concepts emerging from the analysis of sources refer to agent, the resource or the process of reading. In the READ-IT ontology we do not reuse RED or LED terms, but we reuse concepts from the CIDOC CRM family of ontologies and Friend of a Friend (foaf).

6.3.1. Reading Agent

The agent is a human (foaf:Person). The description of the agent is at the time of reading, outlining a specific state in terms of physical, social and cognitive conditions. Thus, Reader represents the states of the agent, aggregation of variable properties describing the agent, a *:descriptionOf* a foaf:Person, see Fig. 7 for an example. Reader is a subclass of cdc:E3_Condition_State and therefore of cdc:E2 Temporal Entity.

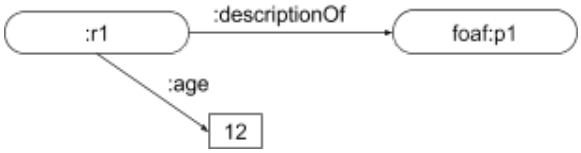

Fig. 7. Reader :r1 of age 12 is a :descriptionOf a person foaf:p1.

As a description of a state of the Person, the Reader is characterised by properties addressing their education, social status, occupation, political stance, religion, age, nationality, gender identity. The question of how to model these concepts are of relevance but fall outside the scope of this specific work. In this regard, the ontology will be revised to include the upcoming module of CIDOC CRM for social documentation, CRMsoc, as soon as it is officially released as well as specific classes developed by READ-IT.

On the other hand, in strict relation with reading, readers often report their reading habits, at the time of reading. Habit is addressed as a class and related to the Reader through the property *:habit*. The characterisation of Habit is one of the open questions that future research on reading should clarify, providing input for further extension of the ontology.

6.3.2. Reading Resource

It is worth to highlight that the content of the testimonies of reading experience do not always provide sufficient information to discriminate between the concepts of work, expression, manifestation and item provided by FRBR. In this regard, the READ-IT concepts of Medium and Content act as an intermediate structure abstracting the FRBRoo implementation of the FRBR concepts.

The Reading_Resource is represented as a disjoint union of Medium and Content, respectively the material and immaterial components of the resource. Medium is represented with the disjoint union of frbr:F3_Manifestation_Product_Type, frbr:F4_Manifestation_Singleton and frbr:F5_Item. Content is represented with the disjoint union of frbr:F1_Work and frbr:F2_Expression. In this frame, an intermediate level property, :providingAccessTo, is also provided to describe the relation between Medium and Content.

The description of experiences could include information about the status of the Medium at the time of reading, e.g. “the book was covered in brown paper”. An Alteration is a :partOf a state of the resource, specifically of the medium. State_of_Medium is a subclass of cdc:E3_Condition_State, and it represents a state of an instance of Medium (:stateOfMedium).

An Alteration can be related to the medium (*:relatedToMedium*) and/or to the content (*:relatedToContent*), e.g. e.g “brown paper covering the book” and a note, Fig. 8.

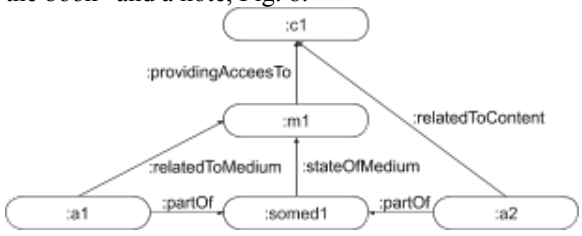

Fig. 8. A State_of_Medium, :somed1 is composed by two instances of Alteration, :a1 and :a2, respectively related to instances of the Medium :m1 and Content :c1, e.g “brown paper covering the book and a note.

6.3.3. Reading Process

Reading, in the sense of human activity, is represented with the concept of Reading_Process. Reading_Process is a subclass of cdc:E2_Temporal_Entity The articulation of a Reading_Process is represented through the concepts of:

- Reading, the full process of reading, from beginning to end, including both active reading and pauses.
- Session, a continuous segment of active reading :partOf Reading
- Experience, a specific moment of active reading :partOf Session

Reading_Process is a subclass of cdc:E7_Activity. Reading, Session and Experience are subclass of Reading_Process.

The Reading_Process is characterised by the literal *engagement* with range "high", "medium" and "low". The engagement represents the level of involvement of the reader in reading, i.e. a spectrum between focus and distraction. Furthermore, *engagement* is specialised as *transportation* to describe the specific involvement with the immaterial component of the reading resource, e.g. story or arguments. See example in Fig. 9.

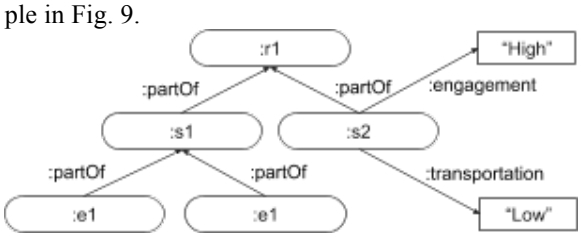

Fig. 9. Reading instance :r1 includes two sessions, :s1 and :s2. The reader reports two experiences in session :s1, while about :s2 provides indication of "Low" :transportation but "High" engagement, e.g. "I was swallowed", "I did not identify with the characters or the story".

Following, Reading_Process implies the existence of at least a Reading_Resource and a Reader. Thus, we define the properties:

- *:involving*, a Reading_Process involves a Reading_Resource
- *:engagedIn*, a Reader is engaged a Reading_Process.

Lastly, a Reading_Process can be a cause of an Alteration of the Medium (*:causeOfAlteration)*, e.g. taking notes, underline. cause of alteration. See example in Fig. 10.

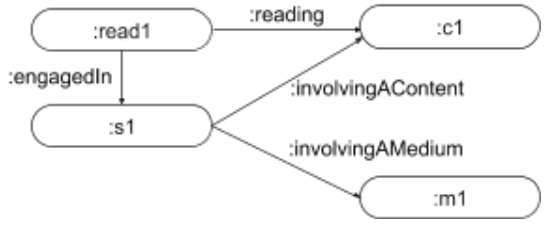

Fig. 10. A reader :read1 is engaged in a session :s1, reading a content :c1. The session :s1 involves the content :c1 and medium :m1.

6.4. Reading Experience

Differently from RED and LED, the experience is represented as a change of the reader's mental state, State_of_Mind, related to the different phases and states of the process of reading, Reading_Process. The core of the reading experience is represented by the relation between a reader's State_of_Mind :effectOf Reading_Process.

State_of_Mind represents a revision of the mental state of the reader. State_of_Mind is a partial description of the new state in terms of which are the new or revised "facts" belonging to the reader's mind. As such, Sate_of_Mind is :partOf Reader (description of the state of the agent).

State_of_Mind is described by the eral *:orientation*, with range:

- *"External"*, description of a change related to the perception of external entities, e.g. objects, activities, people
- *"Internal"*, description of a change related to the self-perception of the reader, e.g. emotions
- *"Undefined"* not applicable or not identifiable

From a temporal perspective, State_of_Mind can occur before (*:precedes*), during (*:coOccuringWith*) of after (*:follows*) a Reading_Process. Specifically, there is a major distinction between the State_of_Mind occurring within the scope of a Reading_Process and the ones occurring outside, before or after.

Following the definition of Reading, Session and Experience, we characterise the Reading_Frame as the union of Reading and Session and the relation

State_of_Mind can have two different relations respect to the Reading_Process: an effect of or a premise to reading. Among effects of reading, we distinguish between State_of_Mind occurring during an active reading and effects occurring after an active reading (in a pause between Sessions or after the end of Reading). In the first case, accordingly with the definition of Reading, Session and Experience, a State_of_Mind is evidence of experiences occurring during the Reading_Process. In the second case, a State_of_Mind is an outcome of a Reading or Session. Thus, we define the properties of:

- *:isEffectOf,* a State_of_Mind occurring during a reading informing about the effects of a Reading_Process
- *:isOutcomeOf,* a State_of_Mind occurring after the end of a Reading_Frame (disjoint union of Reading and Session).

Lastly, we define the property:

- *:isPremiseOf*, a State_of_Mind preceding and informing about a Reading_Frame (disjoint union of Reading and Session).

Summarising, instances of state of mind co-occurring during a Session or Reading should be connected

with an instance of Experience, while following or preceding to instances of Session or Reading, see Fig. 11.

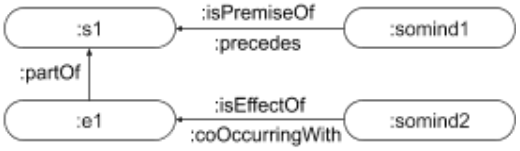

Fig. 11. A State_of_Mind :somind1 is premise of a Session :s1 while :somind2 is an effect of an Experience :e1, part of :s1.

From the analysis of sources, we identified a non-limited list of facets of reader's mind, encoded as subclass of State_of_Mind:

- Self-reflection, reader's self-assessment about the reading and its effects
- Emotion, reader's emotion related to reading process or resource
- Achievement, a deliberation or result related to reader's activities
- Aim, expectations about the reading related to reader's activities
- Remembrance, reader's memories about reading process or resources
- Disposition, reader's stance toward a reading process or resource. Disposition is specialised as follows
  - Aesthetic_Disposition, disposition grounded on aesthetic arguments
  - Ethic_Disposition, disposition grounded on ethics arguments
  - Group_Disposition, disposition grounded on the belonging to a social group or population segment, e.g. teenager, early career, left-party voter
  - Skill_Disposition, disposition grounded on the physical or cognitive skills of the reader, e.g. French proficiency, first-grade education.

The dispositions of the reader are not effects of the reading experience but a type of state of mind oriented toward content, medium or an upcoming reading. A disposition can be related toward a reading resource (:towardInteractingWith) and directed toward an instance of Reading_Frame (:inApproaching), see Fig. 12.

Fig. 12. A reader :read1 reports a state of mind :somind1 outcome of a session :s1, and a disposition :d1 in approaching :s1 toward interacting with a content :c1.

The characterisation of the effects of reading on the reader's mind is one of the major aims of the current research on reading. Therefore, state of mind and its specialisations are yet to be fully described.

6.5. Situation of Reading

Reports of reading experiences often include descriptions of co-occurring events or events as the situation in which the reading occurs. Events can have a direct or indirect relation with reading. In the first case, reading is embedded in a situation while, in the second case, an event is used to make sense of a reading, e.g. by comparison.

In general, we represent events with cdc:E5_Event. In CIDOC CRM, cdc:E5_Event "comprised distinct, delimited and coherent processes and interactions of a material nature, in cultural, social or physical systems involving and affecting instances of E77 Persistent Item in a way characteristic of the kind of process" [12]. This definition addresses the first case, direct relation between event and reading. Indeed, Person and Reading Resource are material entities which are embodied in a social / physical systems, e.g. reading a paper for compiling a survey, borrowing a book from the library, reading to "killing time" during a flight.

Person and Reading Resource are involved in a Reading_Process, specialisation of cdc:E7_Activity and therefore of cdc:E5_Event. We represent the relations between events and reading introducing the property *:situationOf*, with domain and range cdc:E5_Event. For instance, the examination in English literature is *:situationOf* reading the textbook, the degree in modern literature is *:situationOf* all examinations and therefore of all reading related to them, e.g. see Fig. 13.

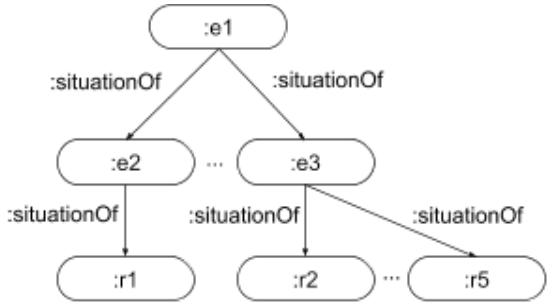

Fig. 13. Event :e1, bachelor in English studies, is a situation of multiple exams, e2, e3, which are situation of multiple reading (of textbooks), r1-r5.

In this frame, the articulation of the Reading_Process in Reading, Session and Experience can be used to represent the implicit hierarchy of reading activity: a Reading instance is the situation of all instances of Session, while an instance of Session acts as the situation of the instances of Experience co-occurring during that instance of Session. In general, as a subclass of E2_Temporal_Entity, temporal properties apply to cdc:E5_Event and therefore events can co-occur, follow or precede other events.

In the case of indirect relations, reading and events are related to each other by the reader. Indeed, there is not a process of interaction of "material nature" outside the reader's mind, but it is the reader's deduction, knowledge, experience that creates the interaction at conceptual level. We address this case introducing a set of comparative properties with domain and range cdc:E5_Event (and therefore cdc:E7_Activity, Reading, Session and Experience):

- *:referredBy*, an event B is being related to an event A
- *:comparableWith*, an event A can be compared with an event B
  - *:betterThan*, an event A is evaluated as veing for some reason better than an event B
  - *:worseThan*, inverse of *:betterThan*

In this frame, instances of reading process can be compared each other, e.g. "despite getting the same score, reading about the industrial revolution was much better than reading about Roman history", see Fig. 14.

Fig. 14. Event e2, exam on Roman history, is comparable with e3, exam on Modern history, while the reading r2, book History of the Industrial Revolution, was a better reading than the reading r2, The Gallic War.

7. Validation

The ontology had been assessed in relation to its internal and external validity [13]. The internal validity had been assessed in terms of a) rigour of the modelling process, and b) adherence to the phenomenon of reading reported in the sources of reading experiences. The rigour of the modelling has partially been addressed in the description of the modelling lifecycle. In the subsection Formal Validity, we report on the tests performed on the RDF and/or OWL encoding of the ontology[13].

It is worth considering that reading is a phenomenon that can only be partially observed, thus the research on reading can rely only on indirect sources documenting reading experiences. In this scenario, the assessment of the adherence of the ontology is assessed as the ability to represent the experience of reading emerging from sources and the research questions of expressed in the use cases [14]. The functional requirements extracted from the sources guided the development of the ontology, as shown by the examples included, while the requirements concerning the research use cases are discussed below (Conceptual Validity).

Lastly, the external validity of the ontology is assessed in relation with the READ-IT project. Specifically, we discuss the non-functional requirements concerning the project architecture (System Requirements) and the research activities (Research Requirements).

7.1. Formal Validity

The technical underpinnings of ontology validation are indicated in the literature as means to automatically test the soundness of an ontology with regard to (1) the underlying logical language used, and

[13] https://github.com/eureadit/reading-experience-ontology

(2) from an engineering-oriented perspective, the ability to answer queries both domain-specific and cross-domain. Both are reported on in the following.

7.1.1. Description Logic Consistency

A basic metric for the validity of any ontology is its consistency with respect to the more expressive description logic (DL) by which it can be interpreted. This includes, at the terminological (TBox) level, whether it defines classes that subsume both the top (owl:Thing) and bottom concept (owl:Nothing), or at the level of facts (ABox), whether it defines individuals that are instances of disjoint classes or violate cardinality restrictions or property domain/range definitions. Assessing consistency requires that disjointness axioms be present, therefore class disjointness was formally defined between all sibling classes in the READ-IT model. The DL consistency of the READ-IT data model was verified by running the HermiT 1.4.3 reasoner[14] over the transitive closure of the imported ontologies (Erlangen CRM, FRBRoo and CWRC).

7.1.2. Expressiveness & Competency Questions

Functional ontology requirements are written in the form of competency questions (CQs) [15].These are defined as questions that the ontology to be built should be able to answer. CQs and their answers play the role of a type of requirement specifications against which the ontology can be evaluated. The idea behind these questions is to ensure that the ontology being developed is committed to the reality being modelled, enough to respond to queries that may be posed to a system that uses the ontology. Thus, CQs also act as a unit test suite for the ontology.

The activity of checking whether the developed ontology is in compliance with a set of ontology requirements is called ontology verification [9]. One approach for performing this activity is (a) to transform (semi-)automatically CQs into SPARQL queries and (b) to check which and how many SPARQL queries obtain a correct response from the ontology.

This activity requires a set of instances covering the whole TBox of the ontology. Such instances can be used to encode the sample response of SPARQL queries. When having this set is not possible, requirements are normally checked in a manual way by means of analysing whether the concepts and relations in the ontology are describing names and verbs in the requirements written as competency questions. This approach is instantiated in Section 7.2.

7.2. Conceptual Validity

The requirements concern specific aspects of the ontology. These requirements emerge from the research questions behind the specific use cases or from a specific type of source. In the following, we address the questions and issues expressed by READ-IT researchers concerning the reader, situation and process of reading, and experience of reading, highlighting when applicable the type of the source.

7.2.1. The Representation of the Reader

- How to represent that a reader was in her youth (letters and diaries)?
- How to represent the changes to the reader's socio-economic status (letters)?
- Is the reader's writing habit within the scope of the model?
- Who are the people who choose to report their emotions (interviews)?
- How can I specify if a reader is an expert?
- In reader psychology, there are theories about links between types of reader and reader response, but the models are built on small studies; e.g. readers from lower socio-educational backgrounds relate book read to their personal experience, but is this the case when using larger samples (interviews)?

These questions concern the characterisation of the status of reader at the time of reading. The ontology provides the class Reader to aggregate the properties concerning age, occupation, nationality, reading habits, gender identity, region, political stand and social status. The collection of the statements of the reader about these aspects of their condition outline a profile of the reader. The characterisation of habit will be addressed in a later stage of the project, as part of the development of case studies and collection of research data in the READ-IT database. About the occupation and social status, UK-RED and LED provide two different characterisations of these concepts. In READ-IT, we did not address concepts concerning the social and personal sphere of the person, but we reuse the upcoming CIDOC CRM module for social structures and social relations, CRMsoc[15], and the

[14] http://www.hermit-reasoner.com/

[15] http://www.cidoc-crm.org/crmsoc/sites/default/files/CRMsoc_20190326.pdf

definitions of previous projects. It is noteworthy that in the study of historical periods and sources, these concepts and the description vary greatly with society values and structures. Thus, it is reasonable to consider an ecosystems of specialist ontologies (for different periods, locations) rather than a specific one.

7.2.2. The Representation of the Situation of Reading

- How do we represent the multiple locations of reading (letters)?
- Is there a link between physical environment and different kinds of reading (interviews)?
- What do you read where and why? Are mobile devices changing the way we read (interviews)?
- Reading aloud: to whom (letters)?
- How do we represent different types of reading, e.g. reading for pleasure or reading for work/study (letters)?
- How I can specify if the reader is reading as part of his/her professional activities?

These questions concern the modalities of reading, e.g. at home on a book, standing on a tram on a smartphone, during a lunch break on an e-reader. Modalities of reading are combinations of place, time, duration and medium. The ontology addresses these aspects through the classes Medium and cdc:E7_Activity. Medium is defined as union of frbr:F3_Manifestation_Product, frbr:F4_Manfestation_Singleton and frbr:F5_Item addressing both physical and digital manifestation of works (e.g. manuscript, eBook) and multiple types of carriers (e.g. printed book, DVDs). The cdc:E7_Activity is a specialisation of cdc:E5_Event and therefore of cdc:E2_Temporal_Entity. As such it addresses temporal aspects, location and participants and properties related to the performance, influences and motivation of the activity. In this frame, the ontology can be used to describe and keep a distinction between 1) the motivation of the activity and the aim of the reader (State of Mind), 2) the objects involved in the activity and the medium used by the reader, and 3) the reading and the activity involving the reading, e.g. class lesson and reading during the class lesson.

- My evidence reports of reading social media, blog posts or other contents which are not books: how should that be encoded in the data model (web contents)?
- In case reading is between multiple contents connected through hyperlink, how do we represent references between content (web contents)?
- How do we encode experiences in which we have information about fragments of text, but information about the title are missing (letters)?

These questions concern the object of reading. Indeed, today, reading is a multi-modality activity (e.g. beginning on a laptop and then switching to a smartphone) that can involve a wide range of types and combination of contents, e.g. posts, web novels, comics, comments, posters. The ontology provides the flexibility to represent complex situations in which the reader's experience is related to multiple reading or reading involves multiple media or content. Furthermore, as previously discussed, the types of content and medium are not limited to printed books or periodicals.

- What if the experience is about an incomplete reading or an attempt to read?
- How to represent something that will be read in the future (letters)?

These questions concern the quality of the interaction between reader and content, specifically about the necessary conditions for considering an interaction as an evidence of a reading experience. A future reading or partial reading implies that a reader is aware of the content that they intend or tried to read. This awareness can be the result of reading the title, the back cover, a review, a summary or upon receipt of a simple suggestion. Aside from suggestions and reviews, all other cases require a direct interaction with the medium and or the content, and a "first impression". A first impression can be considered as a reading experience with a legitimate outcome, e.g. "I wish to read it" or "I don't like it" in which the fragment of the content is the title, the back cover or a blurb. Suggestions (in a written form) and reviews are not considered part of the content, but contents on their own (about another contents).

- How can I quantify the reader's engagement?

This question addresses the evolution of the level of engagement of the reader in the activity of reading. The engagement in reading can be on multiple levels: physical in relation to the interaction with the medium; a cognitive or emotional in relation with the analysis of the content (e.g. arguments, narrative, story or characters). The ontology introduces the class Reading_Process, specialisation of cdc:E7_Activity, to support the characterisation of the engagement. At the current stage, the ontology includes the data properties "engagement" and "transportation" to indicate a level of engagement with, respectively, the process in general or the content.

7.2.3. The Representation of the Experience

- Given different kinds of entries, where do people mention their subjective experience (interviews)?
- In a testimony of reading, how do readers refer to their personal experience, memories, aspirations, identification with character etc. (interviews)?

These questions concern the presentation of the evidences of reading effects in the sources of the reading experience. In this regard, the ontology addresses the encoding of the annotation body, while the W3C Web Annotation Data Model addresses the reference to the portion of source target of the annotation body. The presentation of the experience in the different types of sources can be answered through the study of the fragment of sources annotated as State of Mind and the metadata about the position and structure of these fragments of their target.

- The reader is defining their experience by comparison. Does the model support comparison between experiences?

This question concerns the reader's habit of defining experience by comparison. The ontology provides a set of comparative properties, e.g. better than, and relational properties, e.g. about, to supporting the description of comparison between reading.

- If the reader does not indicate dates but just emotions, how do we represent the experience (letters)?
- What to do when the reader is comparing one book with others, but it is not clear which ones (letters)?
- Not all reading events are transformative [for the reader], reading evidences and not experience only. Which is the minimum set of information required to have a reading experience entry?

These questions concern the minimum set of information required for structuring a representation of reading experience. In general, the lack of information about the effect of reading on the reader is a legitimate piece of information, for instance for the study of how readers report their reading experiences. The ontology can be used to represent an interaction without specific effects or effects of reading without details about the reader, the content or the process.

- In the evidence, the reader is describing emotional aspects of the content and not of their personal reading experience. Is this information in the scope of READ-IT data model (interviews)?

This question points out that the content of a reported experience could be personal or impersonal and states an ambiguity about the content of the experience. In a broader sense, the ontology represents the orientation of a state of mind, e.g. an emotion, which could be oriented toward the self, internal, or other entities and activities. Furthermore, an emotion could be encoded as "emotion" if concerning the response of the reader or as a remembrance if a quoting of the content.

7.3. Ontology and System Requirements

The ontology was evaluated under a new set of requirements emerging from the engagement of research and technology partners.

7.3.1. Types of Sources

The ontology can address annotations on multiple types of sources of reading experience, such as social media, diaries, books, recordings, paintings, video and pictures. The management of annotation is addressed by the concept of "target" of the W3C Web Annotation Data Model (WADM). Specifically, a target can be any resource that can be identified by an IRI (Internationalized Resource Identifier). The description of the target includes information concerning the source (IRI), the style and system of rendering (e.g. software for PDF), selection of the resource (e.g. start and end character counter) and status of the resource at time of annotation (e.g. version). The W3C WADM can represent multiple types of media, individually or as collections.

7.3.2. Research Data

As previously argued, the W3C WADM supports a wide range of different types of sources, while the ontology supports the encoding of the annotations of the different types of reading experience (emerging from the sampling of sources). About the different types of research activities, we rely on the development of different tools, making use of the ontology, designed for supporting specific tasks. In this regard, in READ-IT we consider the following types of tasks.

Crowdsourcing of sources of reading experience, including metadata and licensing through a webtool (as showcased at the SHARP 2019 conference[16]). The new collected sources will be in the scope of the ontology only when annotated

[16] http://www.sharp2019.com/

Manual Annotation of sources by researchers, scholars, students and volunteers. In this regard, a first tool for text annotation is currently being tested. As a result of the first annotation sessions, we identified a subset of concepts of the ontology that will be available through the interfaces of the annotation tool documented in the annotation guide. The data generated through the annotation tool will be integrated through automatic reasoning.

Machine learning and automatic annotation of sources. The data generated from the manual annotation tool and the construction of a training set requires enriching the data with the aim of making the implicit knowledge explicitly encoded. In this regard, the ontology provides an extensive set of properties aiming at the explicit representation of indirect relations, e.g. "reader A reading content B, reading through medium C" can be enriched by stating explicitly that the medium C provides access to content B. Furthermore, the annotation of images and paintings about reading rely on specific visual cues, e.g. a book open or reading to a group of people. In this regard, we introduced the class of State of Medium, and the properties *participants* (to an Event) and *listening* to.

7.3.3. Use of the Ontology

The ontology can support the production of data in the frame of research use cases, and the interoperability of research data beyond the single use cases. The ontology is able to represent the different types of research experiences included in the sources considered in the case studies.

7.3.4. Data Integration

The ontology can support the integration and querying across partial research data. For instance, we consider two applicative scenarios about integration and transversal querying of data about research on an author reading, Example 1 - How to study the reading experiences of Italan poet Ugo Foscolo (1778-1827) in relation to his location, Example 2 - Reading in the Italian Peninsula and during Italian unification.

Example 1 - Studies on Ugo Foscolo's reading

Italian poet Ugo Foscolo lived and worked in several countries during the early 19th century, undergoing changes of social status, political outlook and language during the course of his life.

- *Sources*: letters and critical works
- *Studies:* Foscolo's reading in
  - different countries
  - different languages
  - different socio/economic conditions
  - different political stans

The ontology allows the integration of these different outputs through the concept of Person, the contextualisation of specific reading Events and the analysis of his States of Mind. We can thus query whether at a given time the location of Foscolo's reading experience, his socio-economic situation or the motivation of his reading influenced his evaluations of an author or work and compare his experiences with those of other contemporary readers.

Example 2 - Reading in the Italian Peninsula.

Reading had a central role in the emergence of an Italian national identity during the *19th* century.

- *Sources:* diaries and letters
- *Studies:* Italian readers from multiple locations of the Italian Peninsula during the 19th century

While analysing the diaries and letters of Italian readers from multiple locations, we can query for example whether reading the classics of medieval and Renaissance Italian literature or contemporary political outputs was more common in the various states that composed Italy, and whether the creation of the new Italian state in 1861 changed reading preferences, for example through national school curricula.

7.4. Supporting Research

The management and use of sources in case studies is mostly addressed by the W3C Web Annotation Data Model (WADM). The issues concerning the use of sources are related to their veracity and the reliability of the annotations. Furthermore, researchers point out more conceptual issues related to the aim of the descriptions of reading experience and the social context in which reading and the documenting of the experience are embedded.

- Is there a distinction between fiction and non-fiction content of sources?
- How can we know if annotations are reliable?

These questions concern the value of the information of sources and annotations.

A source can report real or fictional experiences of reading, e.g. a diary is supposed to be a source of real experience while a novel is supposed to be fictional. In general, the evaluation of veracity of the reading experience is grounded on both the type of source and its content. For example, a novel may be fictional but including the real experience of the author, whereas a diary can report third-party comments. In

both cases, the evaluation of the annotator should be reflected on the annotations but not as source metadata.

The reliability of annotations is addressed by the annotation concept of the W3C WADM. Specifically, the W3C WADM addresses the agent creating the annotation (human creator and software generator), the intended purpose and motivation. Quoting the W3C WADM "The creator of the Annotation is also useful for determining the trustworthiness of the Annotation. The software used to create ... the Annotation … is useful for both advertising and debugging issues" [16].

Summarising, the W3C WADM provides the information and references to assess the source (target) and the agent responsible for the annotation (body) but does not provide a structure to report an evaluation as part of the annotation model.

- Can we include anonymised data from reader focus group? (interviews)

The anonymised or pseudo-anonymised transcripts retain the relation between subjects and content. Therefore, transcripts of group meetings can be encoded using the concepts of Person and Reader. For instance, a group including two people each reporting about their readings can be encoded as in Fig. 15.

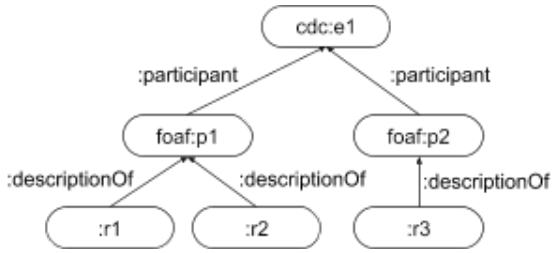

Fig. 15. During an event cdc:e1 (focus group), two people, foaf:p1 and foaf:p2, report about being readers, :r1, :r2 and :r3.

- Can we represent the historical and social context of sources relevant for their correct interpretation and study (diaries)?
- In case the evidence of reading is meant for someone, e.g. the author, how could this information be represented (letters, diaries, periodicals)?
- In case the experience uses different languages, depending on the addressee, how can it be represented (letters)?

These questions concern the relevance of the evaluation of the context in the study of sources. Understanding the context of sources is critical for the annotation and for making use of the annotation in research. For instance, the social context can be the source of emerging patterns about the reading, for instance, specific authors or content subject to censorship or a strong social pressure. This information is outside the scope of the model of reading experience but is partially taken into account by the W3C WADM through the concept of *scope* of the source (IRI), e.g. a Wikipedia page or essay. Indeed, a future development of the project (or a future project) should address the design and development of a repository about the context of sources that could be linked as scope.

8. State of the Art & Related Work

8.1. Validation of Conceptual Models

From the perspective of ontologies in computing, the validation of conceptual models has enjoyed a great amount of research from which a series of key guidelines have emerged, which in turn are implemented across methodologies. Although no holistic validation methodology is in place, there has been an attempt to frame existing criteria, principles and strategies under a semiotic lens and a grouping of criteria into structural, functional and usability-oriented [17]. A recent survey by Degbelo [18] has analysed the merits of ontology validation from an operational as well as a theoretical perspective. Degbelo does not argue that a systematic mapping between validation criteria and strategies should exist - despite still acknowledging bindings between some of these, such as the unsuitability of empirical approaches for the computational efficiency of a model. A strict dependency between validation criteria and development phases does however emerge from the survey.

Some studies introduce the notion of internal and external validity of a model that we have assumed in this work. Guizzardi argues that these may loosely correspond to the "domain appropriateness" and "comprehensibility appropriateness" of language [19], a distinction which Kehagias et al. implement in terms of measuring the cognitive adequacy of an ontology versus its community uptake [20]. Aspects of technical validation, such as ensuring the description logic consistency of the entire ontology network resulting from the model being validated, can be regarded as elements of or preconditions to internal validity. Computational expressiveness, in terms of e.g. which specific description logic family the on-

tology belongs to, is arguably an internal validity factor, if what is being addressed is the decidability or tractability of querying domain data modelled after the ontology.

Further on the internal validity of the ontology, it is worth mentioning the OntoClean methodology for validating the taxonomy relationships of ontologies [21]. The OntoClean approach is based on the systematic analysis of meta-properties of classes, e.g. rigidity, identity and unity, and the consistency of the propagation of the subsumption between classes, e.g. a group of people is also a group. There is limited applicability of OntoClean to this work, as the core concepts and their corresponding taxonomic relationships are imported from CIDOC CRM. Indeed, the overall structure of the reading experience ontology follows CIDOC CRM distinctions between temporal entity and persistent item (disjoint classes), and among the different subtypes of temporal entity: condition state and event or activity.

As for external validity, comprehensibility is one aspect of a more general appropriability of the language which is acquired and transformed by the users who define its pragmatics. Indeed, while comprehensibility is an intrinsic property of the language, its adoption is the result of external factors such as documentation, training and tools. Thus, the evaluation of the ontology can and should take into account its actual use, by means of both machine learning and tools, all contextualised in the domain being defined where possible.

Even under an internal/external lens, however, the partitioning of evaluation criteria along this dimension is not absolute. In particular, the ability of a model to answer competency questions transcends this notion. For one thing, it addresses domain expressiveness as opposed to computational expressiveness. However, CQs can also be considered as criteria for technical validation as much as conceptual, if regarded as having the same role as unit tests for software. Lastly, the ability to translate CQs into formal queries is a crucial factor to an ontology's potential for community adoption, therefore a case can be made for them as external validity factors. The manifold validity of competency questions and their adoption as a tool across several methodologies were influential in the decision to incorporate them in the validation of our model.

With the progressive growth of the Web of Data, methods for validating ontologies against datasets and text corpora, as originally introduced by Brewster et al., have also been gaining continued attention. These methods attempt to respond to a need for objective measures of ontology quality beyond those mandated by the underlying logical framework. The idea behind data-driven ontology evaluation is "to determine how appropriate [an ontology] is for the representation of the knowledge of the domain represented by the texts" [22], although it has been argued that such evaluation metrics would not be exempt from at least temporal or category bias [23]. Although instances of data-driven ontology evaluation methods still form a checkered pattern, some insightful implementations have attempted to compare the model of an ontology to the features extracted from corpora using machine learning and text mining [24]. Though still in its infancy as a proper validation methodology, a corpus-driven approach provides several elements of interest for the domain at hand, which does not inform yet motivates the present study, however this aspect is deferred upon completion of the READ-IT dataset construction.

It should be noted that, although a collection of sources provides an outline of reading phenomenon, it cannot be guaranteed to offer a complete representation. Thus, the ontological commitment of a model should at least express all dynamics emerging from the considered sources, but not be limited to it. Our design choices privileged a focus on addressing false negative examples, rather than constraining the ontology to strictly fit the examples collected. The ability of the ontology to generate scenarios beyond the provided examples was used to verify the validity of the model with the researchers. Indeed, the engagement of researchers highlighted both anti-examples and positive examples, not grounded on the available sources but supported by the current knowledge of the reading phenomenon.

8.2. Related Ontologies

One of the topics discussed in the research on reading experience is related to the history of reading. In this regard, we can refer to the LAWD (Linking Ancient World Data) Ontology[17]. This ontology represents the connection among vocabularies that describe data concerning the ancient world. The LAWD ontology considers a reading as a "word, phrase, or larger chunk of text from a witness (or any observation of variance concerning the text, such as an omission or interpolation)". In addition, the RED ontology is used in datasets that represents the history of reading in Britain from 1450 to 1945. This ontology,

[17] http://lawd.info/ontology

mentioned in Section 3, represents knowledge about reading tastes and habits.

Reading is a language-receptive skill that has a direct connection with listening. For this reason, it is worth recalling the LED ontology (Section 3) as well. This ontology is about listening experience, which is considered as "a documented engagement of an individual in an event of one or more pieces of music being performed".

Reading can imply a direct consequence in the reader related to a severe and temporary mood disturbance, pleasant or painful. This effect is known as emotion and is central to reading experiences [25]. In order to represent emotions, the Emotion Ontology (EMO)[18] was developed. This ontology [26] represents affective phenomena such as emotions, moods, appraisals and subjective feelings. EMO describes the concept "Disposition", which is also included in the READ-IT ontology. This aspect indicates a possible line of combining both ontologies. The different types of emotions are described in-depth in the OCC model [27]. A relevant fragment of the OCC model for the READ-IT ontology is the knowledge related to the emotions concerning consequences of events. This part of OCC model could be aligned to the READ-IT ontology.

9. Conclusions

The development of the READ-IT ontology provides a valuable opportunity to reflect on the role of modelling in the context of research projects, and on how a modelling process can contribute to the creation of new knowledge.

Modelling in the case of the READ-IT ontology moves beyond the encoding of the consolidated domain knowledge of a specific discipline (in this case, history of reading) into facilitating the integration of different disciplinary perspectives and enabling the convergence of knowledge. In particular, the READ-IT ontology transcends the limitations of earlier projects on the history of reading such as UK RED by modelling reading as a human phenomenon rather than a collection of sources. This allows the READ-IT ontology to model the aim of the research process of the project (greater understanding of the role of reading in Europe from 1700 to the present) as well as its starting point (cultural heritage documents that contain evidence of reading). The READ-IT model then enables generative research through the formulation of new, reasoned hypotheses on the experience of reading.

The approach to modelling of the READ-IT ontology supports researchers by establishing a common framework of enquiry. For example, by defining reading agent and reading resource (consisting of medium and content) as two fundamental elements of reading, the model enables Humanities researchers to compare findings and hypotheses even if they are based on data that spans significant temporal and linguistic diversity. The modelling process also highlights questions that are still open for debate, for example those that pertain to the state of mind of the reader, helping to focus current and future research.

With the READ-IT ontology, modelling transcends the encoding of the consolidated domain knowledge of a specific discipline and enables the convergence of knowledge from different disciplinary perspectives. It provides a common framework that can be applied beyond the isolated case studies that constitute the norm in Humanities research to analyse issues that are still under debate and allow the definition of a common object of study.

From a technical perspective, the READ-IT project and the development of its technological ecosystem required rethinking the role of the ontology. In READ-IT, the ontology facilitates the creation of new data sets, addressing limitations of data models used in previous projects. Furthermore, the ontology is used as a reference for the design and development of tools for supporting research activities, such as crowdsourcing reading experience and annotation of sources. The fulfilment of these roles emerged as a precondition for the assessment of the value of the ontology in relation to data and within the project framework.

A challenge worth mentioning concerned the “appropriability” of the ontology by researchers. The complexity of the ontology appeared as an impeding factor for it. Furthermore, the design of a web-based tool for annotation based on the ontology would not have reduced its complexity and could have resulted in discouraging students and volunteers. Indeed, we faced a situation in which the ontology expressiveness was considered correct and appropriate but also an issue.

In this regard, we worked on identifying a subset of concepts of the ontology, a simplified version, which could be easily translated in the annotation tool and be of immediate use for annotators. This solution required a compromise in having a tool generating data on a fragment of the ontology and creat-

[18] http://www.obofoundry.org/ontology/mfoem.html

ing the need to develop an ad hoc reasoner to enrich and complete this type of data source. This solution will enable the creation of new data in a common format without sacrificing the expressiveness of the ontology, which will be of use in the development of other tools and algorithms.

A second challenge concerns the conversion of legacy data from UK-RED (Reading Experience Database). Differently from what was expected, the data collected during the approximately ten years of UK-RED related to reading experience could not be used to validate the ontology. Indeed, the UK-RED data concern concepts and relations almost exclusively related to the reused ontologies (CIDOC-CRM and FOAF). Thus, rather than providing a validation set for READ-IT ontology, UK-RED data provides actually a challenge in terms of preserving the value of a legacy project within a new theoretical and technological framework that is yet to be addressed.

To summarise, the limits of this work are strongly dependent on the challenge it addresses: broadening the scope within the core of research activities; supporting multiple purposes and activities within a yet to be defined technological ecosystem; supporting both disciplinary, multidisciplinary and interdisciplinary research, and current and future research activities; anticipating and facilitating the production of new research data. In this regard, during the next two years of the project, the development of research case studies, technologies and the production of data will provide the opportunity to re-assess the validity of the ontology under the light of the first-hand experience and extend or reframe its most controversial concepts.

Acknowledgments

This work was partially supported by Reading Europe – Advanced Data Investigation Tool (READ-IT) is funded by the JPI Cultural Heritage project under the European Union Horizon 2020 Research and Innovation programme (grant agreement No 699523).

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

This research work has been partially funded by (a) the Research, Development and Innovation Program from the Universidad Politéctinca de Madrid (UPM) (Programa Propio de I+D+i de la UPM) and (b) the project titled System for Evaluating and Adapting Learning Materials to the Easy-to-Read Methodology, approved by the XIX UPM Call for Actions to contribute to the achievement of the Sustainable Development Goals.

The authors would like to thank the READ-IT consortium and associate partners[19] for supporting the development of this work with their invaluable comments and insights.

References

[1] T. Hitchcock, Big Data, Small Data and Meaning, *Historionics Blog*, 2014. http://historyonics.blogspot.co.uk/2014_11_01_archive.html (accessed July 30, 2019).

[2] J. Mussell, Doing and Making, History as Digital Practice, in: T. Weller (Ed.), *History in the Digital Age*, Routledge, London and New York, 2013: pp. 79–94.

[3] J. Laite, The Emmet's Inch: Small History in a Digital Age, *J. Soc. Hist*. (2019). doi:10.1093/jsh/shy118.

[4] J. Flanders, F. Jannidis, Data Modeling, in: *New Companion to Digital Humanities*., Wiley-Blackwell, Malden, MA and Chichester, West Sussex, UK, 2016: pp. 228–237.

[5] L. Putnam, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, *Am. Hist. Rev.* 121 (2016) 377–402. doi:10.1093/ahr/121.2.377.

[6] A. Antonini and L. Lupi, The role of philosophical analysis in the design, in: *Standing on the Shoulders of Giants: Exploring the Intersection of Philosophy and HCI*, CHI 2019, (2019).

[7] A. Adamou, S. Brown, H. Barlow, C. Allocca, M. d'Aquin, Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data, *International Journal on Digital Libraries* 20(1): 61-79 (2019).

[8] M. C. Suárez-Figueroa, A. Gómez-Pérez and M. Fernández-López, The NeOn Methodology framework: A scenario-based methodology for ontology development, *Applied Ontology* 10(2): 107-145 (2015)

[9] M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A. Gangemi (Eds.), Ontology Engineering in a Networked World, Springer, Berlin and Heidelberg, ISBN 978-3-642-24793-4. 2012.

[10] CIDOC-CRM homepage. 2017. http://cidoc-crm.org/ (accessed July 30, 2019).

[11] CRMsoc documentation. http://www.cidoc-crm.org/crmsoc/sites/default/files/CRMsoc_20190326.pdf (accessed July 30, 2019).

[12] CIDOC-CRM documentation v6.2.6. 2019. http://www.cidoc-crm.org/sites/default/files/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf (accessed July 30, 2019).

[13] M. Mitchell and J. Jolley, Research Design Explained (4th Ed) (2001). Harcourt, New York.

[14] F. Vignale, F. Benatti and A. Antonini, Reading in Europe - Challenge and Case Studies of READ-IT, in: *DH 2019 Abstracts*, Utrecht, 2019. https://dev.clariah.nl/files/dh2019/boa/0197.html (accessed July 30, 2019).

[15] M. Grüninger and M.S. Fox, Methodology for the design and evaluation of ontologies, in: *IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[16] W3C Web Annotation Data Model documentation, section 3.3.1 Lifecycle Information. 2017.

[19] https://readit-project.eu/consortium/

https://www.w3.org/TR/annotation-model/#lifecycle-information (accessed July 30, 2019).
[17] A. Gangemi, C. Catenacci, M. Ciaramita and J. Lehmann, A theoretical framework for ontology evaluation and validation, *SWAP 2005*, Vol. 166. 2005.
[18] A. Degbelo, A Snapshot of Ontology Evaluation Criteria and Strategies, in: *Proceedings of the 13th International Conference on Semantic Systems*, ACM, 2017. p. 1-8.
[19] G. Guizzardi, Ontological Foundations for Structural Conceptual Models, 2005.
[20] D. D. Kehagias, I Papadimitriou, J Hois, D Tzovaras, and J Bateman, A methodological approach for ontology evaluation and refinement, in: *The 2nd International Conference of ASK-IT*, 2008, p 1-13.
[21] N. Guarino, Nicola and C. A. Welty, An overview of OntoClean, in: *Handbook on ontologies*, Springer, Berlin, Heidelberg, 2004, p 151-171.
[22] C. Brewster, H. Alani, S. Dasmahapatra and Y. Wilks, Data Driven Ontology Evaluation, 2004.
[23] H. Hlomani and D. A. Stacey, Multiple Dimensions to Data-Driven Ontology Evaluation, in: *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, Springer, Cham, 2014. p. 329-346.
[24] D. Knoell, M. Atzmueller, C. Rieder and K. Scherer, A Scalable Framework for Data-Driven Ontology Evaluation, in: *WM*, 2017, p 97-106.
[25] K. Oatley, Emotions and the story worlds of fiction, in: *Narrative impact: Social and cognitive foundations*, M. C. Green, J. J. Strange, T. C. Brock (eds.), 2002, 39: 69.
[26] J. Hastings, W. Ceusters, B. D. Smith and Kevin Mulligan, Dispositions and Processes in the Emotion Ontology, in: *International Conference on Biomedical Ontology (ICBO)*, 2011.
[27] B. Steunebrink, M. Dastani, Mehdi and J. C. Meyer, The OCC model revisited, In: *Proceedings of the 4th Workshop on Emotion and Computing*, Association for the Advancement of Artificial Intelligence, 2009.