SAPIENZA
UNIVERSITÀ DI ROMA

Visual exploration of digital cultural artifacts

Dipartimento di Ingegneria informatica automatica e gestionale
Antonio Ruberti (DIAG), SAPIENZA – Università di Roma
Ingegneria Informatica (XXXV cycle)

Eleonora Bernasconi
ID number 1726855

Advisor	Co-Advisor
Prof. Massimo Mecella	Prof. Miguel Ceriani

Academic Year December 2022

Visual exploration of digital cultural artifacts
PhD thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: December 2022

Website: http://www.dis.uniroma1.it/users/eleonora%20bernasconi

Author's email: bernasconi@diag.uniroma1.it

«Life is like riding a bicycle.
To keep your balance you must keep moving.»
Albert Einstein

Ringraziamenti

Ritengo che per ottenere qualsiasi grande risultato sia necessario un lavoro di squadra e che non esista vittoria più grande di quella condivisibile con le persone che hanno corso al tuo fianco. Per questo, desidero esprimere la mia gratitudine nei confronti di tutte quelle persone che durante gli ultimi anni hanno: corso con me, tifato per me, creduto in me, condiviso con me difficoltà, gioie e soddisfazioni.

La prima persona che ha avuto su di me un impatto importante riguardo il percorso che mi ha condotta sulla strada della ricerca è stata ***Emanuel****. Emanuel si è dedicato a me con costanza, mettendomi alla prova e insegnandomi cosa significhi fare ricerca.*

La seconda persona è stata ***Francesco****. Dopo che Emanuel aveva fatto nascere in me la passione per l'analisi dei dati, pensai di andare a bussare alle porte del posto dove erano conservate le più grandi quantità di dati riguardanti il nostro paese: l'Istituto Nazionale di Statistica (ISTAT). Quando entrai all'ISTAT capitò, per caso, davanti all'ingresso della portineria, l'incontro con un ricercatore interessato a studenti che avessero la volontà di apprendere l'utilizzo di algoritmi di intelligenza artificiale, da applicare allo sviluppo di soluzioni pratiche, per il proprio dipartimento. Francesco mi ha dato gli strumenti per cavarmela da sola e per navigare con entusiasmo e resilienza in un mondo vastissimo, pieno di insidie e di cose meravigliose: lo sviluppo di applicazioni basate su intelligenza artificiale. Grazie a Francesco ho conosciuto Diego e Monica.*

Diego *è stato il sostegno e la guida nella mia prima esperienza all'estero per una conferenza a Bruxelles.*

Monica *è stata la prima persona a stupirmi per le sue capacità di trasformare "in breve tempo" idee in cose concrete. Monica mi ha dimostrato come sia possibile diventare una donna realizzata, forte, concreta e coraggiosa; mi ha dato inoltre la possibilità e gli strumenti per poter accedere alla mia prima conferenza.*

Dopo Monica ho incontrato ***Massimo****, il relatore di questa tesi, non che fonte inesauribile di opportunità di miglioramento personale, sfide, incoraggiamento, generosità, fermezza, lungimiranza e caparbietà.*

Tutto ciò ha contribuito a cristallizzare in me: la passione per la ricerca; lo studio e la preparazione, per la libertà e l'imparzialità di opinione; il sacrificio e la resilienza, per la libertà personale; il coraggio e la caparbietà, per non guardare solo *al particolare, ma riuscire ad osservare dall'alto le cose in generale.*

Grazie a Massimo ho incontrato ***Miguel****, il correlatore di questa tesi, che mi ha seguita con costanza, in maniera sistematica e generosa, durante tutto il percorso del mio dottorato. Miguel, oltre ad avermi introdotta al mondo dello sviluppo di servizi web e interfacce, mi ha insegnato ad essere una ricercatrice con i principi solidi e fondamentali di: ricercare per conoscere; conoscere per divulgare; divulgare per contribuire.*

Il percorso del dottorato è stato meraviglioso, non solo per il mio miglioramento, a livello personale e professionale, ma anche per i collaboratori e colleghi con cui ho condiviso viaggi, didattica, sviluppo software e progetti.

In particolare ***Francesca****, un'anima buona e devota allo studio, con cui ho svolto ore di lezioni su un corso che considero una delle tante idee geniali del mio relatore Massimo Mecella.*

Alberto*, stimato professionista, programmatore e designer, pieno di conoscenze e generosità con cui passo ore a programmare e a studiare soluzioni innovative.*

Grazie di cuore a Emanuel, Francesco, Diego, Monica, Massimo, Miguel, Francesca e Alberto, persone che hanno contribuito significativamente al meraviglioso percorso di studi, ricerche e lavoro che ho vissuto negli ultimi anni.

L'esperienza insegna che l'effetto di una forza è quello di mettere in movimento il corpo al quale essa viene applicata.

Andrea Elisabetta Giampiero Federica Paola Francesco

Devo loro la mia forza d'animo.

Acknowledgments

I believe that teamwork is required to achieve any great result and that there is no greater victory than the one that can be shared with the people who have raced alongside you. For this, I want to express my gratitude towards all those who, during the last few years, have: raced with me, cheered for me, believed in me, and shared difficulties, joys and satisfactions with me.

Emanuel *was the first person who had a meaningful impact on me regarding the path that led me to the direction of research. Emanuel dedicated himself to me constantly, testing me and teaching me what it means to do research.*

After Emanuel had aroused in me the passion for data analysis, I thought of knocking on the doors of the place where enormous amounts of data are stored concerning our country: the National Institute of Statistics (ISTAT). When I entered the ISTAT, by chance, in front of the entrance to the concierge, I met a researcher.

The researcher was ***Francesco****, who was interested in students who would like to learn the use of artificial intelligence algorithms to be applied to developing practical solutions for his department. Francesco gave me the tools to get by on my own and to navigate with enthusiasm and resilience in a vast world full of pitfalls and beautiful things: the development of applications based on artificial intelligence. Thanks to Francesco, I met Diego and Monica.*

Diego *was the support and guide in my first experience abroad for a conference in Brussels.*

Monica *was the first to amaze me with her ability to transform ideas "quickly" into tangible things. Monica showed me how it is possible to become a fulfilled, strong, concrete and courageous woman. She also gave me the possibility and instruments to access my first conference.*

After Monica, I met ***Massimo****, the advisor of this thesis, an inexhaustible source of opportunities for personal improvement, challenges, encouragement, generosity, firmness, foresight and stubbornness.*

All this contributed to crystallize in me: the passion for research; study and

preparation, for freedom and impartiality of opinion; sacrifice and resilience, for personal freedom; courage and stubbornness, to not only look at the particular, but to be able to observe things in general from above.

Thanks to Massimo, I met ***Miguel****, the co-advisor of this thesis. He followed me consistently, systematically and generously throughout my doctorate. In addition to introducing me to developing web services and interfaces, Miguel taught me to be a researcher with the solid and fundamental principles of researching: researching to know; to know to divulge; disclose to contribute.*

The PhD path was marvelous for my personal and professional improvement and for the collaborators and colleagues with whom I shared travel, teaching, software development and projects.

In particular, ***Francesca****, a good soul and devoted to studying, with whom I taught hours on a course that I consider one of the many brilliant ideas of my advisor Massimo Mecella.*

Alberto*, an esteemed professional, programmer and designer, is full of knowledge and generosity with which I spend hours planning and studying innovative solutions.*

Thanks to Emanuel, Francesco, Diego, Monica, Massimo, Miguel, Francesca and Alberto. They have significantly contributed to the handsome path of studies, research and work I have experienced in recent years.

Experience teaches that the effect of a force is to set the body to which it is applied in motion.

Andrea Elisabetta Giampiero Federica Paola Francesco

I owe my fortitude to them.

Contents

List of Figures

List of Tables

Chapter 1

Extended abstract

Contents

1.1 Research context

Searching and exploring a vast text corpus has often arisen as a human need. Traditionally, the search process is based on manually curated metadata classifying documents by arguments, authors, metadata, etc.

Albeit the metadata that used to be stored in physical cabinets is now stored in databases, the process often remains similar. Although being a decisive paradigm, the maintenance of metadata is costly and becomes progressively more expensive and less reliable with the increase of required detail.

The transition to electronic documents (either created natively as such or digitized) enables the direct text-based search of the content. Text-based search for the full content of documents is a powerful tool. However, it comes with its limitations due to the inherent ambiguity of natural languages and the need for the user to anticipate the actual words used in the content, as the machine cannot capture what the user and the corpus *mean.* This is called the semantic gap. Statistical methods can be successfully used for query expansion, mitigating the issue, but the user has no control of the process.

Semantic enrichment methods, as *named-entity recognition and linking (NERL)* [13, 17], aim to bridging the semantic gap between raw text and concepts, by associating words in the documents with entities in a knowledge base, often a knowledge graph (KG).

NERL successfully enabled users to search and analyze text corpora [16] more effectively. Nevertheless, the navigation of semantic relationships (with their meaning, rather than just as generic connections) between extracted entities has seldom been adopted as a method for the exploration of a corpus; even if it is known that the cognitive processes in library searching are generally more complicated than a single topic-based search [68]. Also, while knowledge extraction methods as NERL are now broadly used by big players in the industry as well as in academic projects, their usage by small to medium size organizations (which often have text corpora, either private or public, that they struggle to manage in a structured way consistently) is still minimal, in part due to the lack of an established standard workflow.

Thus, implementing exploratory semantic search requires various issues, including mapping text to semantic entities, detecting and cleaning inconsistencies in available LOD, ranking algorithms for semantic data and heuristics for recommendations, and appropriate visualizations of complex semantic relationships.

Therefore, the idea proposed in this thesis to address the problems identified concerns a system that, through the use of:

- artificial intelligence techniques, which extract information from unstructured sources (such as text and images);
- semantic technologies that give an unambiguous meaning to the extracted concepts and connect the information with other knowledge bases;
- an interface for explore and query the knowledge graph;
- an interaction paradigm which supports the serendipity effect to discover unexpected things.

Most research in the humanities and related disciplines has focused on small corpora. Data are often processed manually and bound to be used by an institution or one researcher in the context of a project[91, 92]. On the contrary, the building and maintaining institutional systems of organization of knowledge and related datasets in libraries required years of work for a highly skilled and trained workforce. Knowledge graphs in the domain of libraries and digital humanities demonstrate how the application of automatic knowledge extraction and semantic enrichment to large-scale corpora opens up a spectrum of possible new research questions that, until now, were difficult to answer with existing methods.

Exploiting the opportunities in the digital humanities research field poses many methodological and technical challenges.

- Novel user interface and interaction paradigm are needed to support users in viewing, annotating, and systematically analyzing relevant parts of possibly large digitized corpora. Users could express relevance by selecting corresponding concept definitions in knowledge graphs.
- Scalable text-mining and machine-learning techniques are needed to systematically and efficiently analyze and compare the characteristics, contents, and relationships of concepts expressed in knowledge graphs within and across corpora.
- Algorithms are needed that support users in detecting, contextualizing, and analyzing various forms of expressions and associated narrative techniques in corpora spanning an extended period, in which the syntax and semantics may have been subject to constant change.

For these reasons, the need arises to propose:

- the development of tools and scalable techniques for aligning large-scale, multimedia corpora with concepts expressed in knowledge graph;
- support in knowledge exploration through a novel interaction paradigm based on the principle of serendipity that enables users to discover unexpected things.
- the investigation of text mining algorithms that can learn from scholars' annotations and support them in investigating semantic relationships extracted from large corpora;
- the investigation of novel reconciliation mechanisms that ensure that institutional and community-curated knowledge graphs produced in a different context are genuinely inter-operable and do not lead to "competing" data offers;
- validation mechanisms ensure trust in data quality when humans curate data with different levels of expertise and result from automatic processes.

Starting from the background presented, the work discussed in this thesis focused on researching a new interaction paradigm for the management, research and exploration of content automatically extrapolated from a digital library.

This research in a nutshell.

This research aimed at studying the synergy between user interaction, semantic technologies and knowledge exploration tools for digital libraries.
This research has identified open challenges in the Digital Humanities research field related to:

- the user support in viewing, searching and exploring a digital library multimedia content
- the scalability of systems for extracting and exploring a book corpus;
- the difficulty of approaching these tools for users who are not experts in information technology;
- the quality of knowledge extraction;
- the difficulty in maintaining systems based on semantic technologies.

The research competence was demonstrated by studying user interaction with a designed and developed system based on automatic knowledge extraction, semantic technologies and visual search based on the serendipity effect to discover unexpected things. This system is proposed as a possible solution to the open problems identified in the research field of digital humanities.

1.2 Research objectives and contributions

The various research phases related to the study of the state of art, the identification of open challenges and the design development and evaluation process of the proposed solution will be discussed in detail below.

1.2.1 Studying the state-of-the-art

A digital library's knowledge extraction and management are receiving increasing attention in industrial and academic research fields. Thousands of publications use artificial intelligence and semantic technologies tools to reach the research goals. Thus, the first goal that this research aimed to fulfil can be formulated as follows:

Goal 1

Acquire mature knowledge of state-of-the-art techniques and tools to extract and manage the knowledge of a digital library.

A document to be searchable needs to be indexed, either purely based on text content or with the help of tags and information, namely metadata about it, that can be extracted automatically or inserted manually. In the same way, images and other files need metadata. For librarians, most search through images is still based on how someone described its content in a text form, using metadata. For example, Torrossa[1], which is a famous international publishing house, allows the research of their books thanks to detailed manual metadata annotation by domain experts. On the other side, the Yewno[2] platform automatically extracts metadata from books through artificial intelligence techniques based on natural language processing (NLP).

In summary, the first contribution of this research is:

Contribution 1.1

Survey the relevant literature for relevant works, from traditional systems for visual information seeking to tools for semantic enrichment of unstructured text and visualization/exploration of semantic data as KGs, both in the general case and in the specific case of a corpus of books.

1.2.2 Identifying open challenges

More in-depth analysis of the methodologies research that tackled the first objective was focused on mastering the good practices when designing applications for a corpus search and exploration based on semantic technologies. We can summarize this second objective as follows:

Goal 2

Understand how the combination of semantic technologies with a digital library has tackled research challenges in digital humanities research area and identify open challenges in these contexts.

Extensive research was performed on the common problems that the *digital humanist* has recently managed to solve with the help of artificial intelligence and semantic technologies. We find that the knowledge extraction and management systems to disseminate cultural heritage is a particularly active research area in this field. Thus a first contribution is:

[1] `https://www.torrossa.com/it/`

[2] `https://www.yewno.com/`

Contribution 2.1

Systematically reviewed the good practices for designing knowledge extraction and management system for disseminating and enhancing cultural heritage. Open challenges have been identified:

- the inherent ambiguity of natural languages and the need for the user to anticipate the actual words used in the content, as the machine cannot capture what the user and the corpus *mean* and therefore the need to use tools based on semantic technologies;
- the difficulty of approaching semantic technologies based tools for users who are not experts in information technology;
- the difficulty in maintaining systems based on semantic technologies.

From literature emerges like the knowledge graphs are a prominent answer to disseminating cultural heritage challenge. The nature of KG is integrability [25]. This feature allows connecting different cultural domains on the web in the form of linked open data, thus promoting the dissemination of cultural heritage. Furthermore, KG can be explored and interrogated with complex queries favoring the discovery of new knowledge (serendipity). For these reasons, the second contribution to goal number two is the following:

Contribution 2.2

Analyze the applicability and usefulness of a corpus search and exploration paradigm based on the transparent use of knowledge graphs.

1.2.3 Designing and developing an original system

Once that concrete open challenges have been individuated, the pursued task is solving them. The good methodological practices learned in the research path bring ideas and instruments to approach such problems. However, the main contribution of this research was to go beyond. More specifically, tackling the following research goal.

Goal 3

Implement solutions to the open challenges identified in the previous research phase by designing and developing:

- a pipeline for semantic enrichment of textual content;
- a user interface that enables search and exploration of a digital library through visual navigation of a knowledge graph of topics.

The state-of-the-art techniques studied offered many hints to designing proper approaches for these open problems. Specifically, artificial intelligence techniques and semantic technologies were employed. However, novel ideas on how to combine these approaches were necessary to design the proper solutions.
This contribution was the subject of[77, 21, 78]:

Ceriani, M., Bernasconi, E., Mecella, M. *A streamlined pipeline to enable the semantic exploration of a bookstore.* Italian Research Conference on Digital Libraries Springer, Cham. pp. 75-81 Bari (January 2020)

Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T., Capanna, M. C., Di Fazio, C., Marcucci, M., Pender, E., Petriccione, F. M. *ARCA. semantic exploration of a bookstore.* International Conference on Advanced Visual Interfaces Association for Computing Machinery. pp. 1-3 Ischia (September 2020)

Bernasconi, E., Ceriani, M., Mecella, M. *Academic Research Creativity Archive (ARCA).* International Conference on Research Challenges in Information Science. Springer, Cham. pp. 713-714 (May 2021)

From the literature and a study conducted with five researchers belonging to humanities, common behaviors were identified for searching a digital library for content.

The main general supported search behaviors are the following:

- find documents relevant to a specific topic;
- expand or specialize searches by moving through related topics;
- have visibility of available related resources, which could potentially be of interest;

- visually organize the resources found by considering their relationships and properties;
- find topics and documents at the crossing of multiple topics, possibly of different kinds (places, people, time periods, etc.).

For the sake of the analytic approach, the experimentation effort was framed through a set of research goals.

The first contribution to goal three can be expressed as follows:

Contribution 3.1

Relevant research questions to applying KG-based approaches for exploration of text corpora were identified:

- Would users exploring a corpus of text profit from the semantic navigation of the associated KG of topics?
- What kind of user interface would effectively support such a navigation?
- What kind of users, scenarios, and tasks would benefit from this interaction paradigm?
- Does building and maintaining a semantic enrichment and KG creation pipeline necessarily involve high upfront costs and highly skilled developers?

This contribution was the subject of[76]:

Bernasconi, E., Ceriani, M., Mecella, M. *Exploring a Text Corpus via a Knowledge Graph.* 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 91-102 Padua (February 2021)

To answer the above questions, the following hypotheses have been formulated.

Contribution 3.2

Research hypotheses have been identified in response to the research question met.

- Users will be able to effectively explore a text corpus through a KG-based user interface, which offers the following main functions:
  - *a.* finding concepts through text search (among the ones pertinent to the specific domain),
  - *b.* visually navigating the concepts and their relationships, and
  - *c.* showing documents relevant to the selected concept.
- The method, given a corpus of texts in a specific domain, will benefit both users with little knowledge of the domain (by supporting semantically-relevant discovery) and domain experts (by enabling a topic-oriented visual organization of the documents).
- It is feasible to build a ready-to-use complete system, including both semantic enrichment pipeline and web-based front end, which is able, with only some configuration, to be applied to any specific corpus to enable the KG-based exploration.

This contribution was the subject of[76]:

Bernasconi, E., Ceriani, M., Mecella, M. *Exploring a Text Corpus via a Knowledge Graph.* 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 91-102 Padua (February 2021)

While the first two research questions and related hypotheses are relevant for investigating the benefits of the proposed approach for the end users, the last research question and hypothesis investigate the usefulness and portability of such a system to different contexts of use.

Evaluation of the proposed system

Goal 4

Evaluate the proposed solution to improve and measure the level of acceptance (strengths and weaknesses) by the users of the system.

The system has been tested in the context of a specific use case: exploration of the book catalog of medium-size publishing house, specialized in classical antiquity. The anticipated final users of the tool can be roughly classified in two categories:

- domain experts who may adopt a new approach to search and discover resources in the context of their research;
- curious people who want to explore new topics.

The evaluation process lasted two years and was characterized by three phases:

- an evaluation of the extracted data, from the point of view of quality and usefulness, with the help of domain experts;
- a small-scale qualitative user-based evaluation of the tool with a some researchers of the field;
- a larger and richer user-based evaluation of the tool, both on its own and in comparison with other existing solution, which involved both students and researchers of the field.

Contribution 4.1

In two years, three tests were conducted, making it possible to improve the system based on the strengths and weaknesses identified. The evaluations helped to track the results of the efforts applied to the research challenges faced in terms of learnability, usability, effectiveness and user satisfaction concerning the solutions proposed to extract, explore and manage the knowledge of a digital library.
This contribution was the subject of the accepted and under revision journal article[75]:

Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T. *Design, realization and user evaluation of the ARCA system for exploring a digital library.* Journal. International Journal on Digital Libraries Springer. (2022) ***Article under revision***

1.2.4 Extensions of the proposed system

After the proposed system was positively evaluated, it was decided to proceed by developing some applications of potential interest for the domain of digital humanities and digital libraries as extensions of the primary system. In particular, the same interaction paradigm was applied to two different domains concerning the humanities (ancient symbols and ancient places). New functionality has been added downstream of the system that processes the images present in the digital library, recognizing the objects represented, which are inserted into the explorable knowledge domain. Also presented is the extension that allows domain experts to validate the quality of automatically extracted reports.

Application of the interaction paradigm to different domain

Goal 5

Extend the proposed system's functionality by exploiting the automatically extracted information and the information exploration interface.

The availability of a tool such as the one proposed in this thesis would foster collaboration among the researchers in the area, and could attract curious [90] and casual users by easing the diffusion of niche topics like those regarding ancient documentary texts. Offering a pipeline to build a custom KG, can *(i)* introduce a common vocabulary for researchers in the area, *(ii)* share a common understanding of how concepts are related, *(iii)* enable the reuse of domain knowledge, and *(iv)* make domain assumptions explicit. In addition the graphical user interface can be exploited to allow researchers *(i)* to explore the KG, *(ii)* to search and explore relations and connections between resources, *(iii)* to make historical-geographical implications, and *(iv)* to discover new facts about the research field.

Contribution 5.1

Application of the system to the research domain in ancient symbols. Extension of the interface's functionality (display of complex queries) to display semantic connections in common to different resources.
This contribution was the subject of [73]:

Bernasconi, E., Boccuzzi, M., Catarci, T., Ceriani, M., Ghignoli, A., Leotta, F., Ziran, Z. *Exploring the Historical Context of Graphic Symbols: the NOTAE Knowledge Graph and its Visual Interface.* 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 147-154 Padua (February 2021)

Another extension arises from the researchers' need to explore knowledge bases in a cartographic context. This is done by searching for links between the different topics related to a place or a key term, in such a way as to reveal unexpected connections during the exploration of contents and, thus, generating new ideas.

Contribution 5.2

Application of the system to the research domain in ancient geographic locations. Extension of interface functionality to support navigation and display of semantic connections of places on geographic maps.
This contribution was the subject of [74]:

Bernasconi, E., Boccuccia, P., Fabbri, M., Francescangeli, A., Marcucci, R., Mecella, M., Morvillo, A., Tondi, E. *SCIBA-A Prototype of the Computerized Cartographic System of an Archaeological Bibliography.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

Knowledge extraction from images

Another extension of the primary system involved the addition of the automatic extraction of the contents of the images present in a digital library. This extension has allowed the reuse of this information to generate new applications, for example, the semi-automatic creation of book trailers to support storytelling for digital libraries.

Goal 6

Extend knowledge extraction to multimedia content such as book images and create capabilities in the system that leverage this new information.

Multimedia storytelling is an effective and engaging method to convey information in multiple domains. Specifically, book trailers –video advertisements for books– positively influence the desire to learn and the motivation to read. The video trailer generator system supports an expert by gathering relevant crowd-sourced multimedia content, which, arranged as stories, can be used to showcase a book in the form of video clips. Crucially, the expert controls how the content is finally combined and edited rather than offering a fully automated process.

Contribution 6.1

Reuse the knowledge extracted and integrated with other knowledge bases, thanks to semantic technologies, to generate video trailers concerning individual documents' content automatically.
This contribution was the subject of [70, 71, 72]:

Bernasconi, E., Ceriani, M., Mecella, M., De Luzi, F., Sapio, F. *StoryBook. Automatic generation of book trailers.* International Conference on Advanced Visual Interfaces Association for Computing Machinery. pp. 1-3 Frascati (June 2022)

Bernasconi, E., Ceriani, M., De Luzi, F., Sapio, F., Mecella, M. *Storybook: a tool for the semi-automatic creation of book trailers.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

Bernasconi, E., Ceriani, M., De Luzi, F., Di Fazio, C., Marcucci, R., Mecella, M., Sapio, F. *StoryBook-A Storytelling-based Platform for Digital Book Stores.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

Validation of the automatic knowledge extraction

After a critical analysis of the entire information extraction and modeling process, the most significant difficulties arise from problems with the quality of automatically extracted data.

The information extraction workflow was rationalized as much as possible. Given a page of a PDF:

a. this is processed by an OCR;

b. analyzed by the NER to extract names, things, and cities and by the NEL to disambiguate the extracted entities.

c. the entities and metadata related to the inserted document are mapped in RDF on a proprietary KG.

d. the owner KG is connected to external KGs (such as Dbpedia).

The quality of the information extracted depends on these four phases.
Positive and negative feedback was collected during the three evaluations that tested

the quality of the information extracted in each of the four phases. The limitations found, such as OCR errors and disambiguation errors of the extracted concepts, are part of those problems that limit the system potential, which was attested by the users during the evaluation. Consequently to the findings, a strategy for improvement was evaluated founded on the following:

- improving improve as much as possible the OCR, NER, and NEL algorithms [26, 27, 28, 29]
- inserting a human control (human in the loop) to validate the automatic extractions.

At least at the moment, no algorithm can extrapolate information with optimal quality, so the human expert needs to have the last word. For this reason, a proposed solution inserts a layer of human control so that domain experts can validate the automatically extracted results.

Goal 7

Collaboratively improve the quality of automatically extracted information.

Allowing domain experts to collaboratively validate information previously automatically extracted from a digital library is an approach to support the incremental data quality improvement that can be done specifically through the validation of entity linking. Furthermore, rather than seeing just the results of the extraction process, it can be helpful for the domain experts to trace the origin of where the AI recognized a specific entity (i.e. a "snippet" of text or an image).

Contribution 7.1

Integrate into the system the functionality of collaboratively validating in real-time the associations automatically extracted between concepts contained in the books and the books themselves.
This contribution was the subject of [69]:

Bernasconi, E., Ceriani, M., Mecella, M., Morvillo, A. *Automatic Knowledge Extraction from a Digital Library and Collaborative Validation.* International Conference on Theory and Practice of Digital Libraries Springer, Cham. pp. 480–484 Padua (September 2022)

1.2.5 Research projects

Followed are enlist some remarkable projects in which the candidate actively participated during her research and that contributed positively to the development of her technical knowledge and non-technical attitudes as a researcher.

Arca project

The Arca (Academic Research Creativity Archives) project was born from the cooperation between Sapienza University of Rome (operationally the Department of Computer, Control, and Management Engineering Antonio Ruberti and the Department of Sciences of Antiquity), the company Aton Informatica and the historic publishing house L'Erma di Bretschneider. The project was launched in 2019 thanks to the Lazio Region's funding "Creativity 2020", as part of the Operational Program co-financed by the FESR, and aims to develop an innovative digital solution for advanced semantic and bibliographic searches.

The project has favored the development of a system placed in a single market area and with few results produced so far (described in the related work of the paper [76]), that is, it supports users in exploring topics and semantic connections of a catalog of documents. Arca project led to the contribution 6.1 described in section 1.2.3.

Storybook project

Storybook is a research project in the field of digital humanities, which proposes the knowledge extraction and management of information of a digital library to create semi-automatically video trailers of books. Storybook has been originally conceived to meet the needs of "L'Erma di Bretschneider" publishing house that deals with topics related to ancient history and archaeology. From publishers' point of view, promotional trailers respond to a changing market with a high focus on digital and visual media. The goal of a digital presentation of a book is nevertheless broader than selling it and includes providing helpful information to the potential future reader. A publishing house wants to disclose the contents of its digital library not only to experts in the sector but also to interested people attracted by the contents of their books shown on the Web in the form of searching tools or advertising such as video trailers. Numerous researches show that a book trailer fosters the desire to learn and the level of motivation to read [83, 84, 85, 86]. Storybook is a software tool to support the creation of book trailers by collecting and organizing relevant video content. The system users retain control on how to edit and compose the content. The proposed technique aims at semi-automatically building digital trailers that allow the viewers, generically interested in a specialized topic but not expert, to appreciate better the topic, both for their own cultural/professional enrichment and a possible purchase.

Storybook project led to the contribution 6.1 described in section 1.2.4.

Notae project

The project Notae – *Not A writtEn word but graphic symbols. An evidence-based reconstruction of another written world in pragmatic literacy from Late Antiquity to early medieval Europe* –, which started in July 2018, represents the first attempt to investigate the presence of graphic symbols in documentary records as a historical phenomenon from Late Antiquity to early medieval Europe [79]. Notae uses the knowledge graph as a tool to aid the researchers involved in the project and curious users to identify historical and geographical implications.
Notae project led to the contribution 5.1 described in section 1.2.4.

Sciba project

The Sciba (Sistema Cartografico Informatizzato della Bibliografia Archeologica) project envisages the development of an innovative bibliographic search system over an archaeological and historical digital library, on a cartographic basis. The platform, which could later be expanded on a national and international scale, will focus on the Lazio region (Italy), with the support of the "L'Erma di Bretschneider" publishing house, and will allow the visualization of the existing bibliography concerning some selected topographical elements. It will also be possible to expand the search to further related themes, automatically extracted from the system based on the contents and metadata of the integrated bibliographic material. The aim is to develop a search and comparative semantic analysis tool useful for research bodies, museums, municipalities and subjects interested in knowledge and management of the territory, and, at the same time, capable of generating an increase in volume sales with direct employment effects of the publishing chain.
Sciba project led to the contribution 5.2 described in section 1.2.4.

1.3 Publications

1.3.1 Research activities

Parts of the aforementioned work have been published in the following papers:

- **Ceriani, M., Bernasconi, E., Mecella, M.** *A streamlined pipeline to enable the semantic exploration of a bookstore.* Italian Research Conference on Digital Libraries Springer, Cham. pp. 75-81 Bari (January 2020)

- **Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T., Capanna, M. C., Di Fazio, C., Marcucci, M., Pender, E., Petriccione, F. M.** *ARCA. semantic exploration of a bookstore.* International Conference on

Advanced Visual Interfaces Association for Computing Machinery. pp. 1-3 Ischia (September 2020)

- **Bernasconi, E., Ceriani, M., Mecella, M.** *Academic Research Creativity Archive (ARCA).* International Conference on Research Challenges in Information Science. Springer, Cham. pp. 713-714 (May 2021)

- **Bernasconi, E., Ceriani, M., Mecella, M.** *Exploring a Text Corpus via a Knowledge Graph.* 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 91-102 Padua (February 2021)

- **Bernasconi, E., Boccuzzi, M., Catarci, T., Ceriani, M., Ghignoli, A., Leotta, F., Ziran, Z.** *Exploring the Historical Context of Graphic Symbols: the NOTAE Knowledge Graph and its Visual Interface.* 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 147-154 Padua (February 2021)

- **Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T.** *Design, realization and user evaluation of the ARCA system for exploring a digital library.* Journal. International Journal on Digital Libraries Springer. (2022)

- **Bernasconi, E., Boccuzzi, M., Briasco, L., Catarci, T., Ghignoli, A., Leotta, F., Ziran, Z.** *NOTAE: NOT A writtEn word but graphic symbols.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

- **Bernasconi, E., Boccuccia, P., Fabbri, M., Francescangeli, A., Marcucci, R., Mecella, M., Morvillo, A., Tondi, E.** *SCIBA-A Prototype of the Computerized Cartographic System of an Archaeological Bibliography.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

- **Bernasconi, E., Ceriani, M., Mecella, M., De Luzi, F., Sapio, F.** *StoryBook. Automatic generation of book trailers.* International Conference on Advanced Visual Interfaces Association for Computing Machinery. pp. 1-3 Frascati (June 2022)

- **Bernasconi, E., Ceriani, M., De Luzi, F., Sapio, F., Mecella, M.** *Storybook: a tool for the semi-automatic creation of book trailers.*Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

- **Bernasconi, E., Ceriani, M., De Luzi, F., Di Fazio, C., Marcucci, R., Mecella, M., Sapio, F.** *StoryBook-A Storytelling-based Platform for Digital*

Book Stores. Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

- **Bernasconi, E., Ceriani, M., Mecella, M., Morvillo, A.** *Automatic Knowledge Extraction from a Digital Library and Collaborative Validation.* International Conference on Theory and Practice of Digital Libraries Springer, Cham. pp. 480–484 Padua (September 2022)

1.3.2 Other research activities

The results of research activities carried out in parallel with the leading research are listed below:

- **Bernasconi, E., Pugliese, F., Zardetto, D., Scannapieco, M.** *Satellite-Net: Automatic Extraction of Land Cover Indicators from Satellite Imagery by Deep Learning.* ArXiv Preprint. arXiv:1907.09423 . (July 2019)
- **Ziran, Z., Bernasconi, E., Ghignoli, A., Leotta, F., Mecella, M.** *Accurate graphic symbol detection in ancient document digital reproductions.* International Conference on Document Analysis and Recognition Springer, Cham. pp. 147-162 Lausanne (September 2021)
- **Bernasconi, E., De Fausti, F., Pugliese, F., Scannapieco, M., Zardetto, D.** *Automatic extraction of land cover statistics from satellite imagery by deep learning.* Journal. Preprint. Statistical Journal of the IAOS IOS Press. pp. 1-17 (March 2022)
- **Amadori, F., Bardani, M., Bernasconi, E., Cappelletti, F., Catarci, T., Drudi, G., Rossi, M.** *Electrospindle 4.0: Towards Zero Defect Manufacturing of Spindles.* Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

1.4 Thesis outline

Chapter 1 introduce the research problem addressed, its significance in the digital humanities field, and the proposed solution, driven by specific research questions. This serves as the basis for positioning the performed work and summarizing the author's research activities.
Chapter 2 analyzes the background notions about digital humanities. Specifically, it outlines the preliminaries on digital libraries and the research process of digital humanists. Also described the research challenges in the usage of semantic web technologies for digital humanities and presents the employed technique for integrate

and enhance a digital library.

Chapter 3 reports on existent linked data browser tools.

Chapter 4 shows the technical steps enacted to develop the Arca approach as a real implemented tool and presents the results of a multi-step evaluation performed on Arca system.

Chapter 5 presents the extensions of the proposed system.

Chapter 6 concludes the thesis by discussing limitations and future developments. Moreover, it shows results, impacts and benefits addressed by this thesis.

Chapter 2

Background

Contents

2.1 Digital humanities

Digital humanities are a crossover disciplines [22, 23] connecting the social sciences like history, philosophy, archeology, anthropology, linguistic, literature with computer science, media science, and design.

Digital humanities are a two-way approach:

- the systematic use of digital resources in humanities;
- the humanistic analysis of the application of digital resources.

Traditionally, humanities scholars researching specific or multiple literary works are interested in analyzing related texts or text passages. Nevertheless, the digital age has opened possibilities for scholars to enhance their traditional workflows.

Enabled by digitization projects, humanities scholars can nowadays reach a large number of digitized texts through web portals such as Google Books[1].

This shift from reading a single book "on paper" to the possibility of browsing many digital texts is one of the origins and principal pillars of the digital humanities domain, which helps to develop solutions to handle vast amounts of cultural heritage data. The text is the primary data type. Also, tag metadata and images can be managed. In contrast to the traditional methods, the digital humanities pose new research questions on cultural heritage datasets.

2.1.1 Humanities thinking

Computational methods reflect the kinds of research questions that people ask in different disciplinary domains. Even the humanities, which are certainly not monolithic, tend to ask questions that diverge from the sorts of questions prevalent in many fields of the social sciences and the natural sciences.

Digital humanities add something to the conversation around the digital era that is unique. Moreover, that is the attention to exceptions, anomalies, and long tail, rather than the significant patterns that emerge with analytical methods on large data sets. Much of the history of culture is the study of exceptions.
It is masterpieces that changed particular cultural practices, not the prevalent ones. Thus, in the digital humanities fields, there is a split between two types of study approach: a general approach focused on analyzing a vast amount of data versus that tiny subset of data that changed the dominant patterns of the studied domain. Furthermore, those are two very different sets of challenges, and they have substantial social, ethical, and historiographical implications. Computer science can be used to study both, but the methods in question will be different.

There is a straightforward pleasure in the recitation of poetry or the study of a great painting. Furthermore, there is a similar pleasure that one can take in mapping, and graphing, and modeling. Practicing digital humanities for a humanities researcher is an innovative and engaging way of producing knowledge and interpreting the human experience. Humanists can find data that excite them and think about what questions they want that data to help them answer.

Digital methods have changed how people perceive the humanities; there is much information about the humanities that we have yet to discover and uncover. The future of digital humanities looks like a fantastic opportunity to transform how humanists think about the world. In this context, the humanist's role is fundamental, engaging in the very construction of the technologies they will be using.

[1] `https://books.google.it/`

That is the key to the meaning of "digital humanities": The humanists have to change their approach to research on an enormous catalog of documents, and they have to contribute to building the same digital catalog understanding the data formalism, the semantic web, and the under technologies. So, humanists have to become digital humanists to be a part of this transformation process.

2.1.2 History of digital humanities

In 1980 a data entry machine was a room-sized device with a workstation that reads, scans, and turns texts into a readable form. Talking about 1980 suggests that what we now call digital humanities has a long history. That it is not just a product of the digital revolution per se, but instead stretches back in time, not just in 1984, but back to the end of World War II, to the late 1940s, when significant, sometimes room-size systems were used in experimental form to explore humanity's questions.

The earliest examples of what we now call digital humanities were placed under computational humanities, computing and the humanities, humanities informatics.

Thus, the terms have evolved in many decades, but at the core is a conversation between computational techniques, technologies, platforms, and the kinds of questions that characterize the humanities disciplines.

From the late '40s to the 1980s to the present, what we see is a growing portfolio of practices, a growing set of experimental practices that characterize this convergence, but also the collision between the humanities disciplines and computational techniques and tools. So, digital humanities has become the defining label for this field of experimental practice, really with the rise of two key factors. One is the emergence of the internet as the defining public space of our era. Furthermore, the second is the increasing importance of personal computation and personal smart devices as they become not just desktop equipment in offices but fundamental features of our everyday lives.

2.1.3 The library in the digital age

Many people think about the library as being a sort of fixed institution. Like it is a building where people walk inside and get a book. However, digital methods and approaches are straightly changing that.

For many people, the library is still a very physical manifestation of knowledge. It is still a building. It is still a set of materials on a shelf. However, when we think about digital humanities, the concept changes and pushes us towards the digital library and the many ways those traditional library roles have evolved in a new way.

More and more books are published in a digital form or converted to a digital form. Furthermore, that enables new ways to search and new methodologies.

There are also new skill sets that are required within the library. Thus, anyone pursuing library sciences or finds themselves working in libraries ends up with a different set of skills or needs to pursue new skills regularly to keep up with that evolution of digital.

Role of Librarians in the digital age

Librarians have the roles of the acquisition, the description, the curation, and stewardship or preservation of content, whether that is text, book form, or digital content, whether these are images or film or other kinds of formats. The content can include the metadata that describes it, but there are plenty or an increasing number of times when the metadata itself could be a multimedia corpus.

For example, the growth of geospatial data in digital forms and the ability to use GIS, or Geographic Information Systems, to study maps and to study geography and to use geography in other disciplines has become an affluent area for innovation and allows a great exploration.

Other things might include multimedia creation, as that becomes part of the increasing pedagogical shift for how to help students actively learn about their discipline and how to produce information in new ways.

The combination of text and image and other forms such as video has skyrocketed as an essential vehicle for communication.

Any form of human communication becomes an essential vehicle for scholarly communication as well.

So, another area for librarians is visualization and visualization techniques and digital preservation.

How do we store and preserve the integrity of digital objects, not just for current use but for ongoing long-term storage? This question is another challenge where there is an incredible amount of unknown and untested things over time.

2.1.4 Research journey

The research journey of a digital humanist can be built with three phases [24]: materials, processing, and final presentation.

Research begins with materials like texts, images, maps, audio, and any other media files. These materials are analyzed by digital tools to process and extract the answer to research questions. Finally, the results can be presented or archived to pursue other searches.

Behind each of these phases, researchers have much work finding what they expect or discovering new things. If we want to consider it, the time spent in good and exhaustive digital research is much more than the time spent in traditional research due to the vast amount of possibilities and information in the digital world. However, at the same time, for a quick search for information that does not require insights, the digital can instantly answer questions that traditional research cannot do.

Materials

Materials have to be in digital formats to be computationally tractable. This fact requires digital conversion processing if needed (with OCR techniques or digital conversions), the integration of data with the metadata (automatically extracted or not), and other decisions that will have implications for intellectual property, access, sustainability, and use.

Processing

Data processing is the phase where data are modeling with artificial intelligence techniques or digital analysis to build the base of the knowledge extraction from data. This phase is often the black box of the digital humanist's work since much of it can be performed with off-the-shelf tools whose workings may be invisible or incomprehensible to the user. However, contributing to data modeling and having an understanding of these processes allows critical and building engagement. The role of the domain expert, which is a digital humanist, in developing such data modeling is the way to the best research results. Respect specific standards to data modeling allow a project to gain information and data that are curated, updated, and valuable without reinventing existing efforts or focusing on data collection and creation areas peripheral to the investigation.

Presentation

The presentation of the research results often makes use of online platforms such as blogs, WordPress sites, or those designed explicitly for humanists, like Omeka [2] or Scalar[3]. These platforms may be hard to customize and force a humanist research project to conform to a particular argument structure. There is a lack of a tool that allows the humanist to build the research process autonomously, starting from the digitization of documents to the modeling of data in a semantic way, discover new exciting things, visualize the results, and export them to other digital tools or physical supports.

[2] https://www.omeka.net/

[3] https://scalar.me/anvc/

2.2 Semantic web

In 2001 Berners-Lee proposed an extension to the web that would allow people to use computers to process information more effectively.

He called the semantic web solution because this web would add an infrastructure with a defined conceptual meaning - semantics - on top of the syntactic infrastructure of HTML (HyperText Markup Language) for the automatic processing of documents on the web [31, 32].

The semantic web established a new form of the web, in which machines could process data; it was an evolution of the traditional web. The world wide web Consortium's research group on the semantic web defines a common framework (a framework, an area) to share and reuse data, crossing the boundaries between different applications, entities, and communities. The fundamental elements of novelty were:

- links between data and not between documents;
- typed, qualified links.

The semantic web becomes a collaborative movement, led by the world wide web Consortium, an international standardization body. W3C develops technologies that ensure interoperability (specifications, guidelines, software, and applications) to bring the world wide web to its full potential, acting as a forum for information, communications, and shared activities.

The semantic web aims to convert the web characterized by unstructured and semi-structured documents into a web of data, encouraging semantic content in web pages. The semantic web idea can be approached as a linguistic phenomenon that allows the coherent integration of different data. It presents itself as a language for data, invented by humans to communicate fundamentally human information and thoughts. According to well-defined algorithms, the semantic web is conceived to be read and processed not by human readers or listeners but by the computer.

The semantic web, according to Berners-Lee, is “a web of things of the world, described by the data on the web” [33]. The concept is generic, but contains some crucial references:

- the network (graph);
- things (objects related in a meaningful way);
- data (no longer a record, but individual elements, nodes of a network).

Precisely due to the particularities of this type of web, the concept of the semantic web is closely related to the concept of linked data as an effective method and technique for simplifying and homogenizing solutions to identity and interoperability problems with the univocal identification of data in the dialogue between heterogeneous systems. Linked data is a set of techniques that, through shared vocabularies, allows non-human agents to understand content published on the web. This testifies to the progressive transformation of the network into an environment in which all available resources are associated with metatextual information that specifies the semantic context in an accessible format for automatic interpretation and use by part of the search engines.

Therefore, linked data are technology and a set of best practices for publishing data on the web in a readable, interpretable, and usable way by a machine, the meaning of which is explicitly defined utilizing a string consisting of words and markers. The set of corresponding data produces a network: the graph.

The semantic web has the following properties:

- it can contain any data;
- anyone can publish the data;
- the data is self-consistent. That is, it carries its description: if an application encounters data described with an unknown vocabulary, it can dereference the URI (Uniform Resource Identifier) that identifies each term of the vocabulary and arrive at its definition: the self-description mechanism allows the application to resolve the URI that identifies the unknown term to retrieve its definition;
- the entities connected by RDF (Resource Description Framework) [34] links to create a global graph that extends and incorporates different data sources and allows applications to discover new information sources;
- those who publish data on the web are not obliged to use specific vocabularies to represent their data;
- is open: applications can explore and discover new data by following the links in the graph.

Some key concepts are implied in the list of properties of the semantic web:

- the formal and syntactic correctness of a sentence published on the web does not imply equivalent semantic correctness: the statements can be formally correct and, therefore, valid, but not necessarily true; for example, the statement "Steve Jobs was Italian" is formally correct but not true;
- the concept of self-explanatory data, which refers to the concept of dereferenceable URI: the object identified with the URI carries with it a representation of

itself and, therefore, becomes understandable to the agent who encounters it on its exploration path;

- the exploration of the semantic web is unlikely to arise and die within the same set of structured data (dataset): the data published in open mode and connected with qualified relationships create a network that allows them to be used in an unexpected way, not foreseen in the start of the research experience, and allow those inferential processes that constitute the added value of the semantic web: starting from an assertion it is possible to deduce new information, without it being explicitly stated in the data.

2.2.1 Semantic challenges in digital libraries

Digital libraries offer access to large amounts of content in the form of digital documents. Many of them have evolved from traditional libraries and concentrated on making their information sources available to a broader audience, e.g. by scanning journals and books, thereby only taking limited advantage of the benefits modern computing technologies offer. To overcome this bottleneck, research and development for digital libraries include processing, dissemination, storage, search and analysis of all types of digital information. Semantic technologies applied to digital libraries provide an enhancement that helps satisfy users' information needs [35]. Semantic technologies allow for the description of objects and repositories, i.e. the need to establish typical schemes in the form of ontologies, e.g. for the naming of digital objects. The main goal is to enable interoperability, i.e. the ability to consistently and coherently access similar classes of digital objects and services distributed across heterogeneous repositories.

Typical usage scenarios for semantic technologies in digital libraries include:

- user interfaces and human-computer interaction: displaying information, allowing for visualization and navigation of extensive information collections

- user profiling: taking into account the overall information space;

- personalization: balancing between individual and community-based personalization;

- user interaction.

Ontology editors

Ontology editors allow for creating and maintaining ontologies, typically graphically oriented. Familiar to most editors is the ability to create a hierarchy of concepts (such as “Apple is a subconcept of Fruit”) and to model relationships between those

concepts (such as “A person eats an apple”). Many available implementations exist, each having its speciality and different functionalities.
OntoEdit is the most prominent commercial ontology editor[4]. Unlike most other editors, OntoEdit comes with a strong inferencing backbone, Ontobroker, which allows the modelling and use of powerful rules for applications. Numerous extensions, so-called plug-ins, exist to adapt OntoEdit flexibly to different usage scenarios, such as database mapping. Last but not least, full-fledged tool support is provided by the Ontoprise team, which makes it attractive for companies.
Protege[5] is the most well-known academic ontology editor with a long development history. Similar to OntoEdit it is based on a flexible plug-in framework. Numerous plug-ins have been provided, demonstrating possible extensions for typical ontology editors.
KAON[6] is not only an ontology editor but rather an open-source ontology management infrastructure targeted at business applications. It includes a comprehensive tool suite allowing easy ontology creation and management and building ontology-based applications. An essential focus of KAON is on integrating traditional technologies for ontology management and application with those used in business applications, such as relational databases. An example is a PROMPT plug-in, which allows merging two given ontologies into one.

Inference engines

Inference engines allow for the processing of knowledge available on the semantic web. Inference engines deduce new knowledge from already specified knowledge. Two approaches are applicable here:

- general logic-based inference engines
- specialized algorithms (problem-solving methods).

Using the first approach, one can distinguish between different representation languages such as higher-order logic, complete first-order logic, description logic, datalog and logic programming[7]. Inference engines are very flexible and adaptable to different usage scenarios, such as information integration or intelligent advisors, to show the bandwidth of possible scenarios.
Ontobroker[8] is the most prominent and capable commercial inference engine. It is based on frame logic, tightly integrated with the ontology engineering environment

[4] www.ontoprise.com

[5] http://protege.stanford.edu

[6] http://kaon.semanticweb.org

[7] http://semanticweb.org/inference.html

[8] http://www.ontoprise.com

OntoEdit and provides connectors to typical databases. It was already used in numerous industrial and academic projects.
FaCT[36] (http://www.cs.man.ac.uk/horrocks/FaCT/) is one of the most prominent Descriptions of Logics-based inference engines. FaCT (fast classification of terminologies) is a description logic classifier that can also be used for modal logic satisfiability testing. It is based on the tableaux calculus.
KAON2[9] is a new description logics-based inference engine for OWL-DL and OWL-Lite reasoning. The reasoning is implemented with novel algorithms which reduce a SHIQ(D) knowledge base to a disjunctive datalog program.

Data quality

The web is full of incomplete, out-of-date, misleading, biased, inaccurate, contradictory, and otherwise incorrect information, which may arise indeliberately (e.g., due to poor-quality sources, human error, etc.), or deliberately (e.g., vandalism, spam, misinformation, etc.). The web of data is no different, where data quality is an important issue. For example, the Wikidata user "&beer&love" added a date of death for Stan Lee on Wikidata, even though he was still alive (Wikidata revision history for Stan Lee (Q181900), version id. 705733676).

Soon after that, Apple's Siri and Google's Knowledge Panel began reporting Lee's death prematurely before it was flagged and removed (Siri Thinks Stan Lee is Dead - He Isn't (Newsweek)[10]).

While this was likely a deliberate vandalism case, other data quality issues may be more challenging to diagnose and resolve. Data quality can be understood extrinsically in terms of how to fit data are for an external purpose: a dataset that perfectly satisfies one use case might be worthless to another use case. However, data are often published on the web without a principal purpose. It becomes necessary to look at the intrinsic quality of data in terms of dimensions that limit the potential purposes for which it can be used. For example, suppose the majority of a dataset's statements are false. In that case, this is a property of the dataset itself that limits the purposes it can satisfy. Several works have then proposed and analyzed intrinsic dimensions of data quality on the web of data, including:

- the accuracy of coreference (owl:sameAs) links [37, 38];
- conformance of published data to linked data principles [39, 40];
- representativeness [41];

[9] http://kaon2.semanticweb.org
[10] https://www.newsweek.com

- completeness [42];
- etc.

The goal of data quality in the semantic web research field is to assess the quality of available data and improve data published on the web based on assessments. [43] is a survey for a comprehensive list of quality criteria and measures that can be applied for RDF published as linked data.

Link discovery

Key to the success of the current web is the provision of links between relevant documents that help humans and machines discover remote content of relevance to a given search or task. Likewise, links are an essential ingredient of a web of data. Along these lines, the RDF model supports IRIs that can serve both as identifiers and links. At the same time, the linked data principles emphasize the importance of generating links to (and thus receiving links from) remote datasets on the web. Unlike on the current web, links in the context of the web of data can be assigned specific (machine-readable) semantics, which can be very useful for integrating datasets. For example, links using the "owl:sameAs" property can be leveraged to denote coreferent resources in remote datasets. Such links facilitate the discovery of related content elsewhere on the web and the automatic integration of such content based on the semantics of the "owl:sameAs" property, which remarks that the data for both resources can hereafter be merged[37, 38].

Accurately identifying links with specific semantics raises unique challenges like: the quadratic pairwise possibilities for linking between (even just) two datasets raise challenges concerning efficiency; the different ways comparable data can be represented in two datasets present challenges regarding the precision of links produced concerning the stated semantics. A variety of frameworks have been proposed to help address such challenges[44] like:

- Silk [45] relies on handcrafted heuristics;
- LIMES [46] uses distance measures in metric spaces to identify candidates for linking.

Read–write linked data

The web increasingly supports read-write capabilities whereby users can not only read content on webpages but can also write or otherwise contribute content on websites not under their control. These read-write privileges enable the collaborative authoring and curation of web content. Such principles underlie some of the most prominent websites, including wikis, social networks, blogs, etc. The same principles can be applied to the web of data, where read-write privileges can enable users to collaborate on content [57]. Like on the current web, arbitrary users should not have complete write access over data on remote websites, but rather policies, access control, etc, should govern the types of contributions they make.

Also, on the current web, depending on the website, users may lose jurisdiction over their content. Some problems are:

- users may not be able to control how their content is used;
- by whom the content is used;
- how long content remains available;
- user may lose the ability to export or adapt the content for their purposes.

Implementing read-write capabilities on the web of data furnishes an opportunity to address such issues. Along these lines, Berners-Lee presented Read–Write linked data[11] as a way for remote users to collaboratively edit and curate data on the web in a decentralized manner. The main crossroads towards this vision was the standardization of the linked data Platform (LDP)[12]. It lays the foundations for read-write linked data applications. Several LDP implementations are available [58, 59, 60] for such applications. However, while the LDP specification outlines key protocols, other practical details still need to be addressed regarding decentralized authentication, access control, etc.

Annotation tools

Annotation tools[13]

allow for adding semantic markup to documents or, more generally, to resources. The great challenge here is to automate the annotation task as much as possible to reduce the burden of manual annotation for large-scale resources.

Annotea[14] is a "Live Early Adoption and Demonstration" project enhancing the W3C collaboration environment with shared annotations. By annotations, we mean comments, notes, explanations, or other types of external remarks that can be attached to any web document or a selected part of the document without actually

[11] T. Berners-Lee. Read-Write Linked Data. W3C Design Issues, Aug. 2009.https://www.w3.org/DesignIssues/ReadWriteLinkedData.html.

[12] S. Speicher, J. Arwe, and A. Malhotra. linked data Platform 1.0. W3C Recom-mendation, Feb. 2015.https://www.w3.org/TR/ldp

[13] http://annotation.semanticweb.org

[14] http://www.w3.org/2001/Annotea

needing to touch the document. When the user gets the document, he or she can also load the annotations attached to it from a selected annotation server or several servers and see what his peer group thinks.
OntoMat-Annotizer[15] is currently the most prominent annotation tool. It is based on a full-fledged annotation framework which is already being extended to support semi-automatic annotations of documents and databases.
KIM[16] provides a knowledge and information management infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content.

Query interfaces

SPARQL[17] provides an expressive language for posing queries against the web of data. The challenge is that relatively few web users are likely to be familiar with its syntax. Hence, research on query interfaces usable by non-experts is crucial for the web of data to reach a broader audience. Ideally, such user interfaces should not be locked to a certain functionality but should instead allow users to generate queries in an intuitive but general manner. That said, there tends to be a trade-off between the expressiveness of a query interface and its usability: more complex interfaces allow for generating more complex queries but are often more challenging to use.

Various works have explored this trade-off between expressivity and usability in query interfaces. Classifying the query interfaces for the web of data into three main categories:

1. The first category regards question-answering systems [47, 48]. These systems accept a query in natural language and convert it to a structured query (e.g., SPARQL) or try to answer it directly over the available data. These systems are easy to use but currently toil to answer more complex questions.

2. The second category is facetted search [49, 50, 51]. The user begins with a list of results (e.g. given for a keyword search) and can then select facets upon which to refine the results shown iteratively. Though intuitive, such systems typically only allow for expressing a particular structure's acyclic queries and returning a list of entities as results.

3. The third category is query builders [53, 54, 55, 56]. These systems assist users in constructing graph patterns that can be evaluated over the data, generally

[15] `http://annotation.semanticweb.org/ontomat`

[16] `http://www.ontotext.com/kim`

[17] `https://www.w3.org/TR/rdf-sparql-query`

using visualizations, auto-completion, etc. Although such systems allow users to pose complex queries over a dataset, they are commonly more challenging to use and require some knowledge of how the data are structured.

Additional advances regard:

- improving the accuracy of question-answering systems;
- the usability of query builders;
- the inclusion of reasoning capabilities into such interfaces [50].

Knowledge graphs

Knowledge graphs have been gaining more and more attention in industry and academia alike, where they serve as a graph-structured substrate of knowledge within a particular organization [61, 62]. It is possible to distinguish two types of knowledge graphs:

- open knowledge graphs are published online for the public good (e.g., DBpedia [63], Wikidata [57], etc.);
- enterprise knowledge graphs are typically internal to a company and used for commercial use-cases (e.g., Google Knowledge Graph, etc. [62]).

Enterprise knowledge graphs often benefit from sources on the web of data, pulling content from Wikidata [57], Schema.org [162], etc. In research, knowledge graphs intersect several areas, particularly:

- graph databases;
- knowledge representation;
- logic;
- machine learning;
- information extraction;
- graph algorithms;
- etc [61].

This confluence of areas gives rise to novel research questions regarding:

- how to combine machine learning and knowledge representation;
- how to leverage graph algorithms for information extraction;

- etc.

Currently, some of the main trends in this area are:

- knowledge graph embeddings, which aim to learn numeric representations of knowledge graphs [64];
- knowledge graph refinement, which aims to assess and improve the quality of knowledge graphs along various dimensions [64];
- exploring applications for knowledge graphs in domain-specific settings such as to promote tourism [67]
- enhance cultural heritage [65]
- combat human trafficking [66]
- many others [61].

2.3 Knowledge graph

The knowledge graph is a visual representation of a given domain's interlinked entities (e.g., people, places, events), which captures the information about these entities and forges connections between them. A knowledge graph is a suitable method of representing entities published as Linked Open Data (LOD) resources. It might give a more straightforward solution for tracing different entities' connections.

As a data structure, they underpin a digital information system, support users in information discovery and retrieval, and are helpful for navigation and visualization purposes. Within the libraries and humanities domain, knowledge graphs are generally embedded in knowledge organization systems, which have a century-old convention and have experienced a digital transformation with the advent of the web and linked data.

In recent years, knowledge graph technologies established a solid position in the enterprise world, serving as a central element in the organizational data management infrastructure. Knowledge graphs are becoming both the repository for organization-wide master data (ontological schema and static reference knowledge) and the integration hub for various legacy data sources: e.g., relational databases or data streams.

Metadata and concept definitions are now constructing an interconnected and decentralized global knowledge network that can be curated and increased by community-driven editorial processes. In the future, knowledge graphs could be conveyances for formalizing and linking findings and insights derived from the analysis of possibly large-scale corpora in the libraries and digital humanities domain.

2.3.1 Knowledge bases of humanities

The development and curation of domain-specific knowledge structures have traditionally been an essential element of humanities disciplines. In many cases, these structures emerge as an implied research output, e.g., when studying the interrelation between actors and events in a specific historical background. In other cases, developing a knowledge organization system as such or its components is the research effort's main objective, like developing domain taxonomies, gazetteers or prosopographic (dictionaries of people or groups of people).

Initiatives such as the Tabula Imperii Romani[18], the Tabula Imperii Byzantine, the Prosopography of the Byzantine World[19] or the Treasury of Lives[20] have been concerned exclusively, sometimes over a significant timespan, and long before the digital transformation, with the curation of authoritative data on places or people within their domain.
Other efforts have focused on translating existing information to digital: linked data. For example:

Pleiades[21] [97] contains data about geographical places (e.g., cities) relevant for the study of ancient literature and history.
Papyri.info[22] is a research engine integrating several different ancient document databases.
MANTIS is the semantically enriched database of the American Numismatic Society[23], a research institute devoted to study coins from all periods and cultures.
Open Context[24] collects, among other resources, archaeological reports.
Trismegistos [99] is a metadata platform[25] for the study of texts from the Ancient World. It contains data about ancient documents, people, and places.
EDH[26] is a search tool for Latin epigraphic data.

Developing classification schemes and authority information in the humanities requires structure, control, and a common vocabulary to facilitate collaboration.

Scholarly humanities research differs insofar as there is much more specificity to particular or niche domains.

[18] `https://tir-for.iec.cat/`

[19] `https://tib.oeaw.ac.at/`

[20] `https://treasuryoflives.org/`

[21] see `https://pleiades.stoa.org/`

[22] see `http://www.papyri.info/`

[23] see `http://numismatics.org/search`

[24] see `https://opencontext.org/`

[25] see `https://www.trismegistos.org/`

[26] Epigraphic Database Heidelberg - see `https://edh-www.adw.uni-heidelberg.de/`

A global knowledge graph of the humanities seems an unachievable goal because: the heterogeneous nature of research results; the interpretative quality of humanities research; often work is organized around the efforts of a single individual or small group, financed via time-limited grants.

Nevertheless, the need to publish data under open licenses and build connections between datasets based on shared value vocabularies and metadata element sets is increasingly perceived in the community as key for: enabling re-use; for the transparency of scholarly approaches; promoting results.

This trend is mirrored in the promotion of linked data as a theme at digital humanities events, conferences, and curricula, as much as in the emergence of community initiatives dedicated to establishing common interlinking standards and approaches.

Such initiatives advocate the idea of cross-domain linkage employing shared name authority files and recommendations on metadata element use. Crucial to this effort is, on the one hand, generic knowledge graphs like DBpedia [27] and Wikidata[28].

Linked datasets of libraries or museums have been gaining traction as an interconnecting spine through which community-specific datasets can build outbound links to contribute to a global graph (e.g. the Virtual International Authority File[29]; the Getty Thesaurus of Geographic Names[30]; the Getty Art and Architecture Thesaurus[31]).

Most studies in the humanities and related disciplines have focused on relatively small corpora. Data was manually curated and often bound to an institution or the scope of an individual researcher's project. In contrast, the constitution and maintenance of institutional knowledge organization systems and related datasets in libraries required years of work for a highly skilled and trained workforce. Large-scale knowledge graphs in the libraries and digital humanities domain demonstrate how the application of quantitative analysis to large-scale corpora opens up a spectrum of possible new research questions that, up until now, were hard to answer with existing methods.

Exploiting the opportunities of quantitative analysis methods poses many methodological and technical challenges:

- Novel serial analysis methods and tools are needed to support scholars in viewing, annotating, and systematically analyzing relevant parts of possibly large digitized corpora. Scholars could express relevance by selecting corresponding concept definitions in knowledge graphs.

[27] `https://www.dbpedia.org`

[28] `https://www.wikidata.org`

[29] `https://viaf.org`

[30] `https://www.getty.edu/research/tools/vocabularies/tgn/`

[31] `https://www.getty.edu/research/tools/vocabularies/aat/`

- Scalable text-mining and machine-learning techniques are needed to systematically and efficiently analyze and compare the characteristics, contents, and relationships of concepts expressed in knowledge graphs within and across corpora.
- Algorithms are needed that support scholars in detecting, contextualizing, and analyzing various forms of expressions and associated narrative techniques in corpora spanning an extended period, in which the syntax and semantics may have been subject to constant change.

For these reasons, the need arises to propose:

- the development of tools and scalable techniques for aligning large-scale, multimedia corpora with concepts expressed in knowledge graph
- the investigation of text mining algorithms that can learn from scholars' annotations and support them in investigating semantic relationships extracted from large corpora;
- the investigation of novel reconciliation mechanisms that ensure that institutional and community-curated knowledge graphs produced in a different context are genuinely inter-operable and do not lead to "competing" data offers;
- validation mechanisms ensure trust in data quality when humans curate data with different levels of expertise and result from automatic processes.

There is a clear need for collaborative research among humanities, computer science, and library information science researchers. This need also results in novel mixed qualitative and quantitative methods for analyzing large-scale digitized corpora relevant to the humanities and related disciplines.

Chapter 3

Related work

Contents

3.1 Introduction

In this chapter, relevant literature is surveyed for relevant works starting from tools for semantic data management, to traditional systems for visual information seeking, visualization/exploration of semantic data as KGs and collaborative annotation, both in the general case and in the specific case of a corpus of books.

Figure 3.1 lists the characteristics and functionalities of tools that deal with the management and representation of semantic data.

The existing literature on interfaces and tools for linked data was analyzed to select the tools, and the forty-four tools most representative of the identified categories were chosen (as a practical example).

In figures 3.2 and 3.3 are listed the tools to which the categories they belong have been assigned.

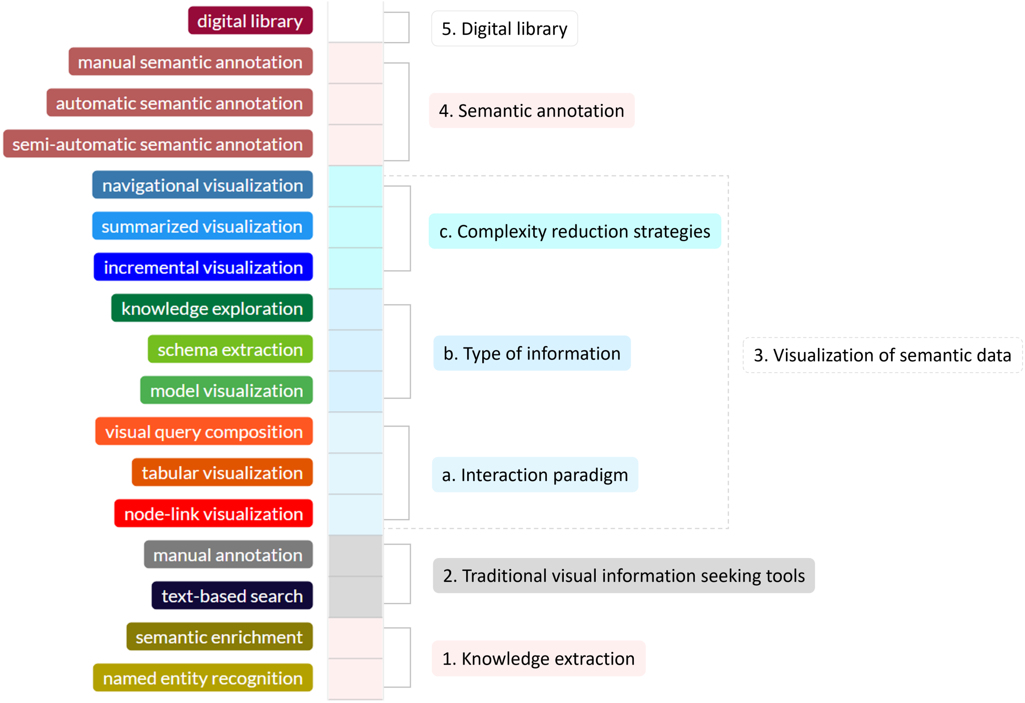
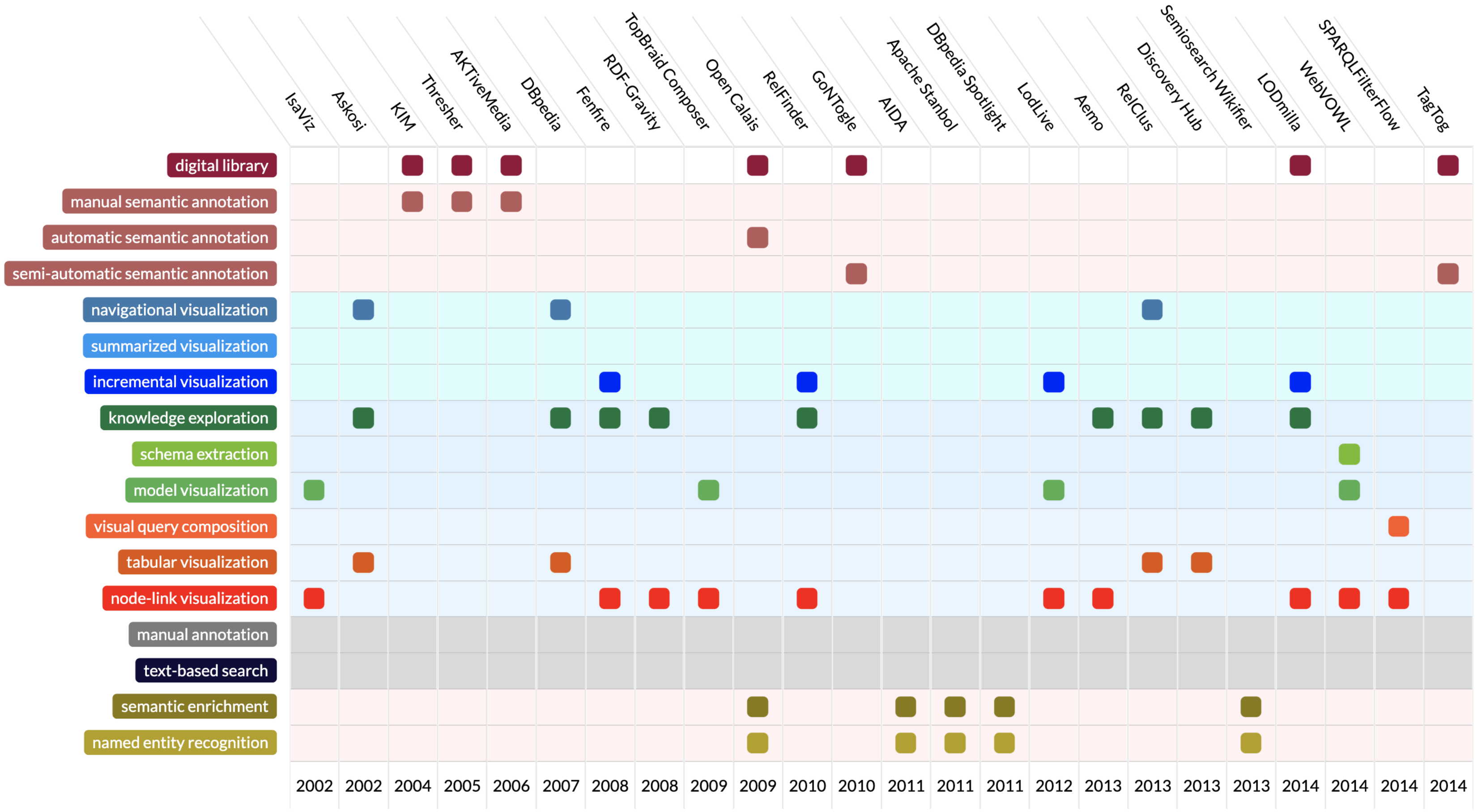


Figure 3.1. Classification categories

Some categories aim to identify the visualization and interaction paradigm used, such as graph-based visualization and incremental visualization. In contrast, others are concerned with background characteristics, such as semantic enrichment and knowledge extraction.

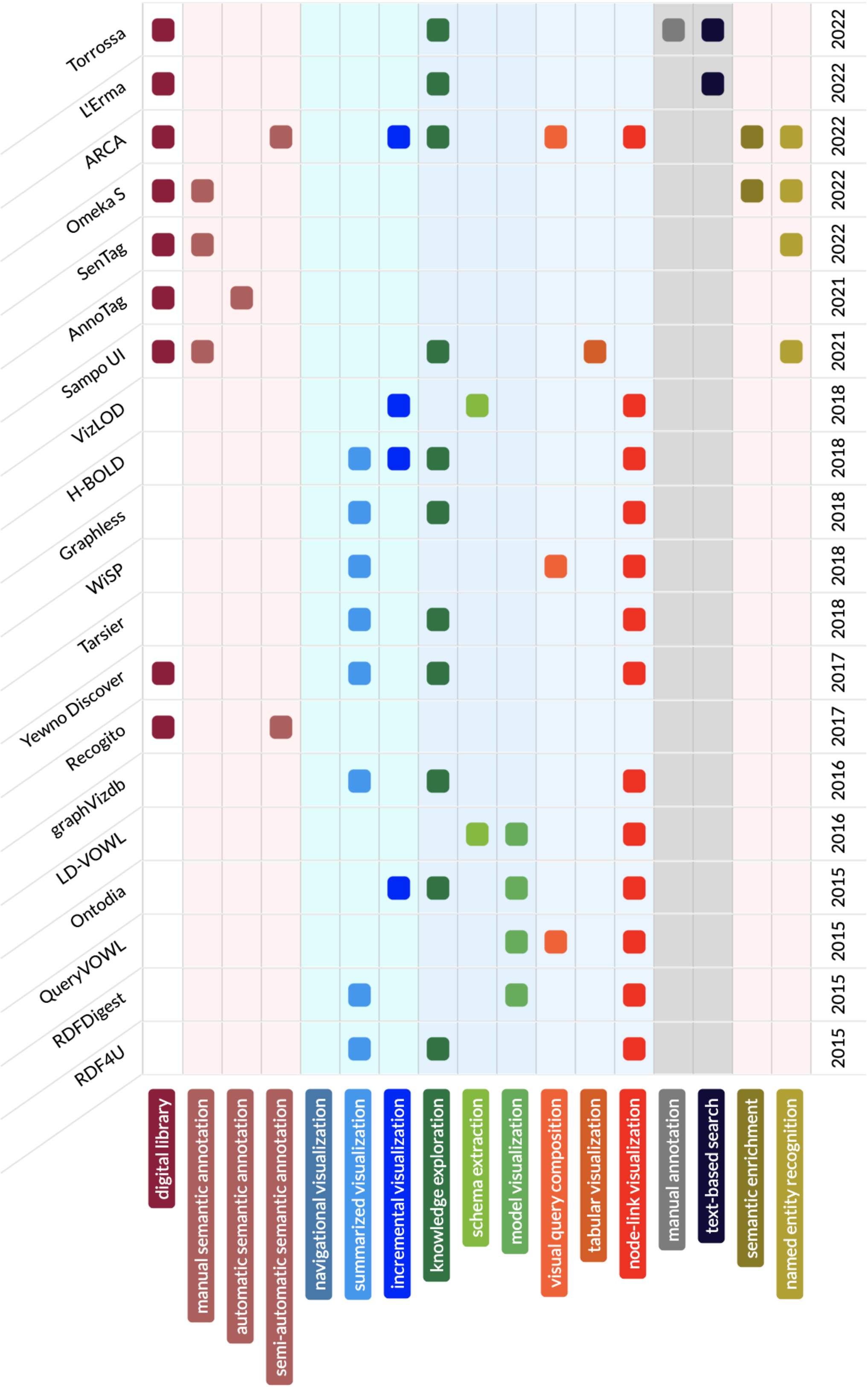
3.2 Knowledge extraction

Knowledge Extraction aims to lift an unstructured or semi-structured corpus into an output described using a knowledge representation formalism. Thus Knowledge Extraction can be seen as Information Extraction but with a stronger focus on using knowledge representation techniques to model outputs. Information extraction from unstructured sources and semantic enrichment are crucial tasks for managing semantic data. Some tools that manage these activities are described below.



	IsaViz	Askosi	KIM	Thresher	AKTiveMedia	DBpedia	Fenfire	RDF-Gravity	TopBraid Composer	Open Calais	RelFinder	GoNTogle	AIDA	Apache Stanbol	DBpedia Spotlight	LodLive	Aemo	RelClus	Discovery Hub	Semiosearch Wikifier	LODmilla	WebVOWL	SPARQLFilterFlow	TagTog
digital library			■	■	■					■		■									■			■
manual semantic annotation			■	■	■																			
automatic semantic annotation										■														
semi-automatic semantic annotation												■												■
navigational visualization		■				■												■						
summarized visualization																								
incremental visualization							■				■					■					■			
knowledge exploration		■				■	■	■			■						■	■	■		■			
schema extraction																						■		
model visualization	■								■							■						■		
visual query composition																							■	
tabular visualization		■				■												■	■					
node-link visualization	■						■	■	■		■					■	■				■	■	■	
manual annotation																								
text-based search																								
semantic enrichment										■			■	■	■					■				
named entity recognition										■			■	■	■					■				
	2002	2002	2004	2005	2006	2007	2008	2008	2009	2009	2010	2010	2011	2011	2011	2012	2013	2013	2013	2013	2014	2014	2014	2014

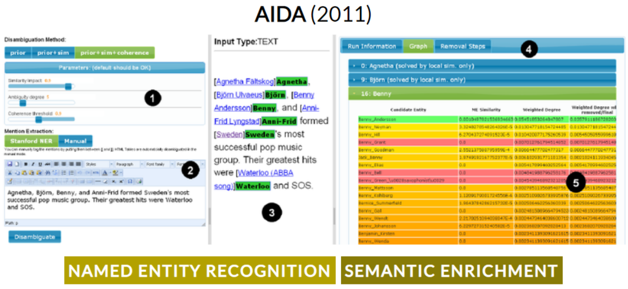
Figure 3.2. Semantic web tools - first part



		digital library	manual semantic annotation	automatic semantic annotation	semi-automatic semantic annotation	navigational visualization	summarized visualization	incremental visualization	knowledge exploration	schema extraction	model visualization	visual query composition	tabular visualization	node-link visualization	manual annotation	text-based search	semantic enrichment	named entity recognition
Torrossa	2022	■							■						■	■		
L'Erma	2022	■							■							■		
ARCA	2022	■			■			■	■			■		■			■	■
Omeka S	2022	■	■														■	■
SenTag	2022	■	■															■
AnnoTag	2021	■		■														
Sampo UI	2021	■	■						■				■					■
VizLOD	2018							■		■				■				
H-BOLD	2018						■	■	■					■				
Graphless	2018						■		■					■				
WiSP	2018						■					■		■				
Tarsier	2018						■		■					■				
Yewno Discover	2017	■					■		■					■				
Recogito	2017	■			■													
graphVizdb	2016						■		■					■				
LD-VOWL	2016									■	■			■				
Ontodia	2015							■	■		■			■				
QueryVOWL	2015										■	■		■				
RDFDigest	2015						■				■			■				
RDF4U	2015						■		■					■				

Figure 3.3. Semantic web tools - second part

AIDA [114] (fig. 3.4) is a framework and online tool for named entity recognition and resolution. Given a natural-language text or a web table, it maps mentions of ambiguous names onto canonical entities (e.g., individual people or places) registered in the YAGO2 knowledge base, also used to provide sense tagging. AIDA can be configured for the algorithm to be applied (prior probability, keyphrase similarity, coherence). It is available as a demo web application or a Java RMI web service.



You can manually tag the mentions by putting them between [[and]]. HTML Tables are automatically disambiguated in the manual mode.

Agnetha, Björn, Benny, and Anni-Frid formed Sweden's most successful pop music group. Their greatest hits were Waterloo and SOS.

[Agnetha Fältskog]Agnetha, [Björn Ulvaeus]Björn, [Benny Andersson]Benny, and [Anni-Frid Lyngstad]Anni-Frid formed [Sweden]Sweden's most successful pop music group. Their greatest hits were [Waterloo (ABBA song)]Waterloo and SOS.

0: Agnetha (solved by local sim. only)

9: Björn (solved by local sim. only)

16: Benny

Candidate Entity	ME Similarity	Weighted Degree	Weighted Degree w/ removed/final
Benny_Andersson	0.0010497821536934663	0.05451853264947807	0.0357011186728[illegible]
Benny_Neyman	3.324827854826432E-5	0.013047718154724485	0.0130477181547244[illegible]
Benny_Hill	6.270437274091523E-5	0.010420877175260539	0.0054505[illegible]
Benny_Grant	0.0	0.007012761794514852	0.0070127617945148[illegible]
Benny_Goodman	2.552137580795595E-4	0.006644777024773317	0.0066447770247731[illegible]
Jack_Benny	1.9749192167752377E-5	0.006102931771101354	0.0021024113934[illegible]
Benny_Elias	0.0	0.005417899460252564	0.0054178994602525[illegible]
Benny_Bell	0.0	0.004841988796258176	[illegible]
Benny_Green_\u0028saxophonist\u0029	0.0	0.00454394689232120[illegible]	[illegible]
Benny_Mattsson	0.0	0.002785113568548756	[illegible]
Benny_Kohlberg	1.1209179081724556E-4	0.0025109026733995876	0.002510902673399[illegible]
Bernice_Summerfield	1.9643784286215732E-5	0.002506462256360339	0.0025064622563603[illegible]
Benny_Gall	0.0	0.0024815089664794523	0.0024815089664794[illegible]
Benny_Wendt	2.2170251094039847E-4	0.0024473464039860071	0.0024473464039860[illegible]
Benny_Johansson	6.229727315240581E-5	0.002368207092028413	0.0023682070920284[illegible]
Benjamin_Kirsten	0.0	0.0023411393091621615	0.002341139309162[illegible]
Benny_Wanda	0.0	0.0023411393091621615	0.002341139309162[illegible]

Figure 3.4. Aida

Apache Stanbol[1] is an Open Source HTTP service meant to help Content Management System developers semi-automatically enhance unstructured content with semantic annotations to link documents with related entities and topics. Current enhancers include RDF encoding results from multilingual named entity recognition and resolution, sense tagging regarding DBpedia and GeoNames, text span grounding, confidence, and related images. It is available as a demo web application, REST service, or downloadable.

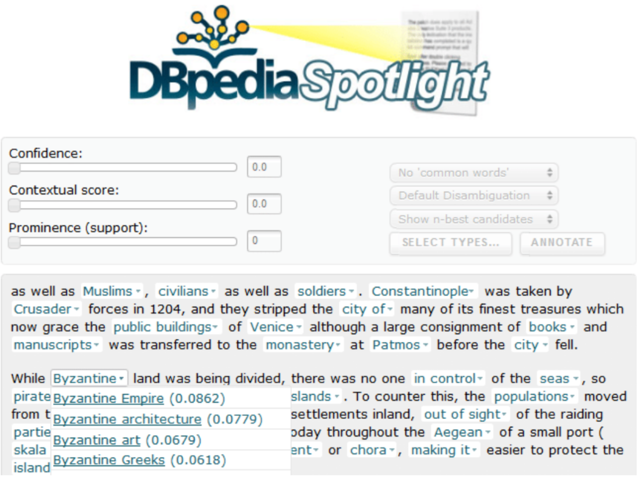
DBpedia Spotlight [113] (fig. 3.5) is a tool for automatically annotating mentions of DBpedia resources in text. It is available as a demo web application, REST service, or downloadable.

Open Calais [52] (fig. 3.4) is a KE tool that extracts named entities with sense tags, facts and events. It is available as a web application and as a web service. It has been used via the web application for homogeneity with the other tools. We have also tried the Open Calais TopBraid Composer plugin, which produces an RDF file automatically. The RDF schemata used by Open Calais have mixed semantics and have to be refactored to be used as a standard output relevant to the domain addressed by the text.

Semiosearch Wikifier [145] (fig. 3.7) resolves arbitrary named entities or terms (i.e., individuals or concepts) on DBpedia entities by integrating several components: a named entity recognizer, a semiotically informed index of Wikipedia pages, and

[1] `https://stanbol.apache.org`

DBpedia Spotlight (2011)



Confidence: 0.0
Contextual score: 0.0
Prominence (support): 0
No 'common words'
Default Disambiguation
Show n-best candidates

as well as Muslims, civilians as well as soldiers. Constantinople was taken by Crusader forces in 1204, and they stripped the city of many of its finest treasures which now grace the public buildings of Venice although a large consignment of books and manuscripts was transferred to the monastery at Patmos before the city fell.

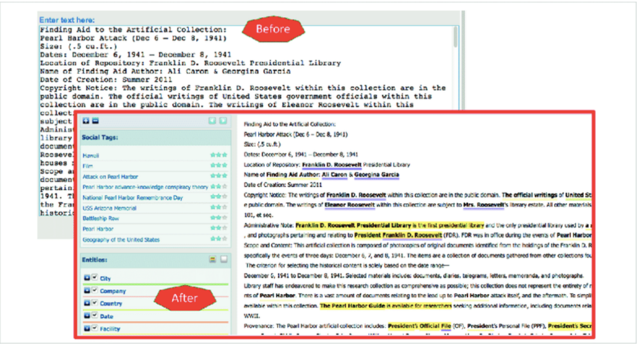
While Byzantine land was being divided, there was no one in control of the seas, so pirate slands. To counter this, the populations moved from t settlements inland, out of sight of the raiding partie oday throughout the Aegean of a small port (skala ent or chora, making it easier to protect the island

Byzantine Empire (0.0862)
Byzantine architecture (0.0779)
Byzantine art (0.0679)
Byzantine Greeks (0.0618)

NAMED ENTITY RECOGNITION SEMANTIC ENRICHMENT

Figure 3.5. DBpedia Spotlight

Open Calais (2009)



Enter text here:

```
Finding Aid to the Artificial Collection:
Pearl Harbor Attack (Dec 6 – Dec 8, 1941)
Size: (.5 cu.ft.)
Dates: December 6, 1941 – December 8, 1941
Location of Repository: Franklin D. Roosevelt Presidential Library
Name of Finding Aid Author: Ali Caron & Georgina Garcia
Date of Creation: Summer 2011
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the
public domain. The official writings of United States government officials within this
collection are in the public domain. The writings of Eleanor Roosevelt within this
collect
subject
Administ
library
document
Roosevel
houses
Scope a
document
pertain
1941. T
the Fra
histori
```

Social Tags:
Hawaii ★★★
Film ★★★
Attack on Pearl Harbor ★★★
Pearl Harbor advance-knowledge conspiracy theory ★★☆
National Pearl Harbor Remembrance Day ★★☆
USS Arizona Memorial ★★☆
Battleship Row ★★☆
Pearl Harbor ★★☆
Geography of the United States ★★☆

Entities:
☑ City
☑ Company
☑ Country
☑ Date
☑ Facility

Finding Aid to the Artificial Collection:
Pearl Harbor Attack (Dec 6 – Dec 8, 1941)
Size: (.5 cu.ft.)
Dates: December 6, 1941 – December 8, 1941
Location of Repository: Franklin D. Roosevelt Presidential Library
Name of Finding Aid Author: Ali Caron & Georgina Garcia
Date of Creation: Summer 2011
Copyright Notice: The writings of Franklin D. Roosevelt within this collection are in the public domain. The official writings of United St
e public domain. The writings of Eleanor Roosevelt within this collection are subject to Mrs. Roosevelt's literary estate. All other materials
101, et seq.
Administrative Note: Franklin D. Roosevelt Presidential Library is the first presidential library and the only presidential library used by a
, and photographs pertaining and relating to President Franklin D. Roosevelt (FDR). FDR was in office during the events of Pearl Harbor
Scope and Content: This artificial collection is composed of photocopies of original documents identified from the holdings of the Franklin D. R
specifically the events of three days: December 6, 7, and 8, 1941. The items are a collection of documents gathered from other collections fo
The criterion for selecting the historical content is solely based on the date range—
December 6, 1941 to December 8, 1941. Selected materials include: documents, diaries, telegrams, letters, memoranda, and photographs.
Library staff has endeavored to make this research collection as comprehensive as possible; this collection does not represent the entirety of
nts of Pearl Harbor. There is a vast amount of documents relating to the lead up to Pearl Harbor attack itself, and the aftermath. To simpli
available within this collection. The Pearl Harbor Guide is available for researchers seeking additional information, including documents rela
WWII.
Provenance: The Pearl Harbor artificial collection includes: President's Official File (OF), President's Personal File (PPF), President's Secr

NAMED ENTITY RECOGNITION SEMANTIC ENRICHMENT AUTOMATIC SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.6. Open Calais

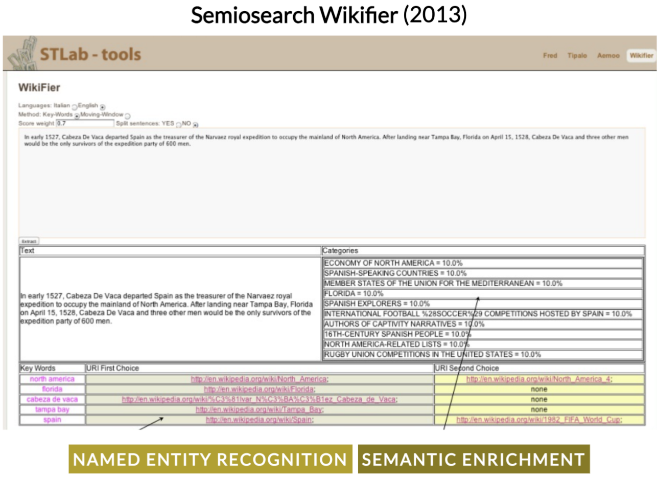
matching and heuristic strategies. It is available as a demo web application.

There has been recently much research on how to attach semantics to unstructured data [16, 87], through processes like NERL.

The **GLOBDEF system** [14, 88] works with pluggable enhancement modules, which are dynamically activated to create on-the-fly pipelines for data enhancement.

Both tools provide interesting paradigms to build a flexible pipeline for semantic enrichment.

In comparison with the system proposed in this thesis, they do not provide directly



Semiosearch Wikifier (2013)

WikiFier

Languages: Italian ○ English ●
Method: Key-Words ● Moving-Window ○
Score weight 0.7 Split sentences: YES ○ NO ●

In early 1527, Cabeza De Vaca departed Spain as the treasurer of the Narvaez royal expedition to occupy the mainland of North America. After landing near Tampa Bay, Florida on April 15, 1528, Cabeza De Vaca and three other men would be the only survivors of the expedition party of 600 men.

Text	Categories
In early 1527, Cabeza De Vaca departed Spain as the treasurer of the Narvaez royal expedition to occupy the mainland of North America. After landing near Tampa Bay, Florida on April 15, 1528, Cabeza De Vaca and three other men would be the only survivors of the expedition party of 600 men.	ECONOMY OF NORTH AMERICA = 10.0% SPANISH-SPEAKING COUNTRIES = 10.0% MEMBER STATES OF THE UNION FOR THE MEDITERRANEAN = 10.0% FLORIDA = 10.0% SPANISH EXPLORERS = 10.0% INTERNATIONAL FOOTBALL %28SOCCER%29 COMPETITIONS HOSTED BY SPAIN = 10.0% AUTHORS OF CAPTIVITY NARRATIVES = 10.0% 16TH-CENTURY SPANISH PEOPLE = 10.0% NORTH AMERICA-RELATED LISTS = 10.0% RUGBY UNION COMPETITIONS IN THE UNITED STATES = 10.0%

Key Words	URI First Choice	URI Second Choice
north america	http://en.wikipedia.org/wiki/North_America;	http://en.wikipedia.org/wiki/North_America_4;
florida	http://en.wikipedia.org/wiki/Florida;	none
cabeza de vaca	http://en.wikipedia.org/wiki/%C3%81lvar_N%C3%BA%C3%B1ez_Cabeza_de_Vaca;	none
tampa bay	http://en.wikipedia.org/wiki/Tampa_Bay;	none
spain	http://en.wikipedia.org/wiki/Spain;	http://en.wikipedia.org/wiki/1982_FIFA_World_Cup;

NAMED ENTITY RECOGNITION SEMANTIC ENRICHMENT

Figure 3.7. Semiosearch Wikifier

a front end to use the semantic information for information retrieval, which is crucial for the stated purposes and scientific questions of the present work. Furthermore, while GLOBDEF and Stanbol testify about the interest in this kind of solution, neither of them is actively maintained, the former being stuck in prototype status while the latter has been retired, and thus not practically usable to test the stated hypotheses.

3.3 Digital humanities knowledge bases

The world of digital humanities offers several different digital services and Web applications providing information about ancient documents, historical facts and geographical entities, many of which are based and/or can be downloaded as semantically enriched datasets following the LOD principles. The ones that can potentially be of interest for the domain of this thesis are listed below.

Pleiades[2] [97] contains data about geographical places (e.g., cities) relevant for the study of ancient literature and history. Pleiades allows the download of their data in JSON, CSV, and RDF/TTL format from a Github repository.

Papyri.info[3] is a research engine integrating several different ancient document databases. It models relationship between documents from different sources using RDF triples.

MANTIS is the semantically enriched database of the American Numismatic

[2] see `https://pleiades.stoa.org/`

[3] see `http://www.papyri.info/`

Society[4], a research institute devoted to study coins from all periods and cultures. In MANTIS, each record can be exported as JSON-LD, Turtle RDF/XML, and many others.

Open Context[5] collects, among other resources, archaeological reports. The data can be exported in tabular (CSV) or Geo-JSON form.

Trismegistos [99] is a metadata platform[6] for the study of texts from the Ancient World. It contains data about ancient documents, people, and places. Trismegistos does not currently provide a data export feature, but it does share its data with Papyri.info in the form of periodic database dumps.

EDH[7] is a search tool for Latin epigraphic data. It provides a data dump including RDF (inscriptions including prosopography) and GeoJSON formats.

All of them have embraced the LOD philosophy as a mean to connect their data to the broader digital cultural heritage infrastructure. Unfortunately, all of these efforts are jeopardized by the lack of a common vocabulary and the lack of integration. In addition to these services, a wide set of websites is available, containing digital reproductions of ancient original documents. Many of these websites are provided by university departments, libraries and archives, and do not provide standard ways to access data. An approach similar to ours, with a narrower target, has been proposed in [98] for studying people (producing prosopography) who lived in Ptolemaic Egypt, based on the Trismegistos data.

Finally, Pelagios [100] is an international consortium that uses the LOD approach and the Pleiades gazetteer to join up a variety of online resources that refer to places in the ancient world.

3.4 Traditional visual information seeking tools

There has been a large amount of work in the literature about visual information seeking [18, 3, 10]. The first attempts to create a visual search interface, have been done in the early 1990s [2], where some researchers had applied direct manipulation principles to search interfaces, creating what they called dynamic queries [1].

These are visual query systems, often based on the query-by-example paradigm [20]: search interfaces where users can manipulate sliders and other graphical controls to change search parameters. The results of those changes are immediately displayed to them in some visualization.

As an example relevant to our case, in the **site of the publishing house**

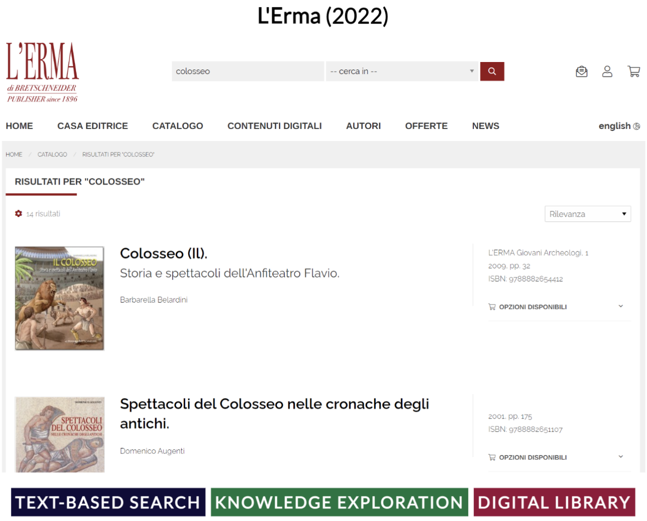
[4] see `http://numismatics.org/search`

[5] see `https://opencontext.org/`

[6] see `https://www.trismegistos.org/`

[7] Epigraphic Database Heidelberg - see `https://edh-www.adw.uni-heidelberg.de/`

"L'Erma di Bretschneider"[8] (fig. 3.8) there is a traditional book search system that allows searching by keywords contained in book titles and by categories. For conciseness we will call this search system *Lerma*.



RISULTATI PER "COLOSSEO"

14 risultati

Colosseo (Il).
Storia e spettacoli dell'Anfiteatro Flavio.
Barbarella Belardini

L'ERMA Giovani Archeologi, 1
2009, pp. 32
ISBN: 9788882654412

Spettacoli del Colosseo nelle cronache degli antichi.
Domenico Augenti

2001, pp. 175
ISBN: 9788882651107

Figure 3.8. L'Erma di Bretschneider

Torrossa[9] (fig. 3.9) is the digital search platform of "Casalini Libri" to which about 180 publishers, mainly Italian and Spanish ones, adhere with their contents. Torrossa allows an advanced search by metadata and by words contained in the books. A limit of these systems, for unstructured information like books, is that exploring and filtering by basic metadata (i.e., author, title, etc.) can be useful, but it is often insufficient.

Therefore, semantic entities assigned to documents extend the capabilities of traditional keyword-based search by:

- proper semantic facetted browsing to filter search results;
- extension of the query string with related entities and keywords;
- recommendations of related documents and further search suggestions by following cross-connections.

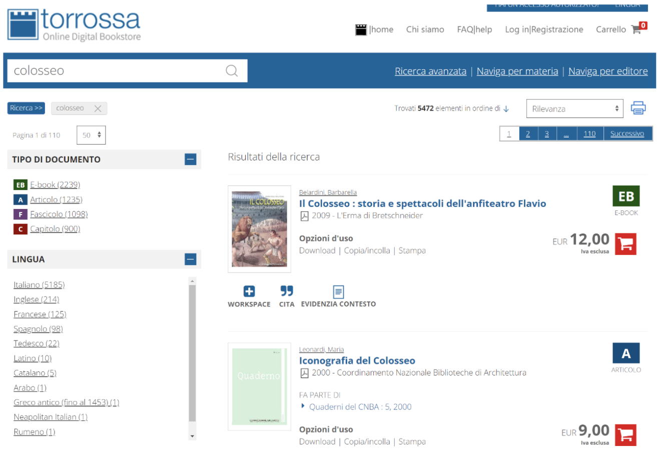
3.5 Visualization of semantic data

In Section 3.2 tools for visualization of semantic data have been surveyed.

[8] http://www.lerma.it/

[9] https://www.torrossa.com/

Torrossa (2022)



Trovati **5472** elementi in ordine di ↓

Pagina 1 di 110

TIPO DI DOCUMENTO

- EB E-book (2239)
- A Articolo (1235)
- F Fascicolo (1098)
- C Capitolo (900)

LINGUA

- Italiano (5185)
- Inglese (214)
- Francese (125)
- Spagnolo (98)
- Tedesco (22)
- Latino (10)
- Catalano (5)
- Arabo (1)
- Greco antico (fino al 1453) (1)
- Neapolitan Italian (1)
- Rumeno (1)

Risultati della ricerca

Belardini, Barbarella
Il Colosseo : storia e spettacoli dell'anfiteatro Flavio
2009 - L'Erma di Bretschneider
EB E-BOOK
Opzioni d'uso
Download | Copia/incolla | Stampa
EUR **12,00** Iva esclusa

Leonardi, Maria
Iconografia del Colosseo
2000 - Coordinamento Nazionale Biblioteche di Architettura
A ARTICOLO
FA PARTE DI
Quaderni del CNBA : 5, 2000
Opzioni d'uso
Download | Copia/incolla | Stampa
EUR **9,00** Iva esclusa

TEXT-BASED SEARCH MANUAL ANNOTATION KNOWLEDGE EXPLORATION DIGITAL LIBRARY

Figure 3.9. Torrossa

The extracted semantics can be extremely useful for exploring a corpus of documents, but they are not fixed and homogeneous like a set of predefined metadata. Therefore, data models and visual user interfaces need to deal with these complex and heterogeneous data. The semantic web [4] and linked data [6] efforts deal with data modelling, integration, and interaction of this kind of data on the web. These efforts lately contributed to the emergence of KGs to organize complex data-sets integrating multiple sources [19, 9].

Many user interfaces for visualization and exploration of KGs exist, and new ones are being developed every year, especially using semantic web and linked data technologies [15, 8, 11, 5].

3.5.1 By interaction paradigm

Several tools have been developed offering basic interactive operations that allow users to visually explore linked data graphs provided by either a data file or SPARQL endpoint.

This subsection lists the tools for viewing linked data classification by interaction paradigm. The interaction paradigm concerns using tools and processes to produce a visual representation of the data that can be explored and analyzed directly within the visualization. Different interaction paradigms support different data-driven insights methodologies.

We have identified three types of interaction paradigms:

- **Tabular**: In interfaces with a tabular interaction paradigm, information about a single resource is shown in one visualization. Views focus on a table or multiple tables, which show specific properties linked to the asset, such as media files such as photos, descriptions, or links to other linked assets.
- **Node-link**: In the node-link paradigm, resources are represented by nodes or boxes, while triples are represented by arcs that connect the resources. The node-link view can be static or dynamic (the latter allowing interaction).
- **Visual query composition**: The visual query paradigm comprises user interfaces that the user to perform advanced queries without necessarily having technical knowledge about the RDF model and the SPARQL language.

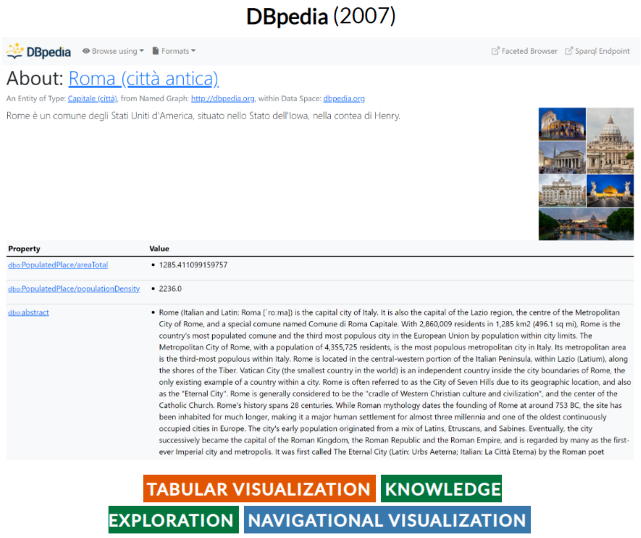
Tabular visualization tools

DBpedia[10] (fig. 3.10) is a project aiming to extract structured content from the information created in the Wikipedia project. DBpedia is presented in tabular format as a list of triples. The resource page allows users to view the list of all the connections (all the triples in which the resource acts as a subject/object), including inbound and outbound arcs. The user's interaction with this interface provides that all the predicates and objects, when they are resources, can be clicked. That is, they are links (written in blue and underlined) that allow you to jump from one page to another, all resources, therefore, have links that allow exploration of related resources.

Node-Link visualization tools

Aemoo [117] (fig. 3.11) aims to provide additional information than the mere content that can be found in a SPARQL endpoint, it provides exploratory search over the Web. The tool receives information from the DBPedia endpoint and improves that content with information gathered over Wikipedia, Google News, and Twitter. Its primary focus is to provide information about the resource building a bridge between the Semantic Web and the traditional Web. Aemoo uses Wikipedia as its primary source of information. It collects all Wikilinks connected to that resource and divides them into "set nodes". Each "set node" contains resources of the same class. Finally, it explores Google News and Twitter searching for more information connected to the subject. All the set nodes are displayed on a graph where the core node is the resource of interest. Hovering an entity will show contextual information about the relation between the subject and the hovered element; generally, that information is a sentence found on Wikipedia. Aemoo approach is based on knowledge

[10] `https://www.dbpedia.org`



About: Roma (città antica)

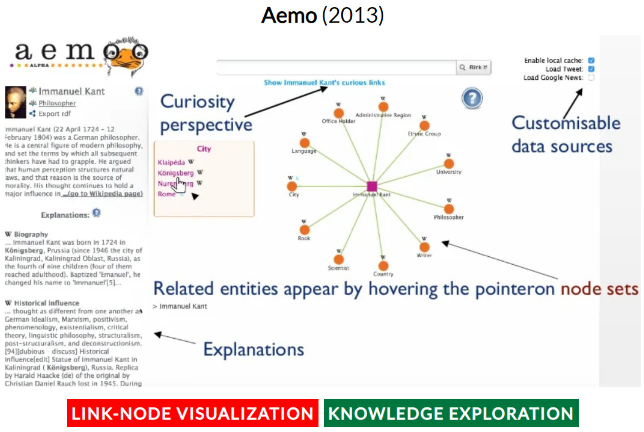
An Entity of Type: Capitale (città), from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Rome è un comune degli Stati Uniti d'America, situato nello Stato dell'Iowa, nella contea di Henry.

Property	Value
dbo:PopulatedPlace/areaTotal	• 1285.411099159757
dbo:PopulatedPlace/populationDensity	• 2236.0
dbo:abstract	• Rome (Italian and Latin: Roma [ˈroːma]) is the capital city of Italy. It is also the capital of the Lazio region, the centre of the Metropolitan City of Rome, and a special comune named Comune di Roma Capitale. With 2,860,009 residents in 1,285 km2 (496.1 sq mi), Rome is the country's most populated comune and the third most populous city in the European Union by population within city limits. The Metropolitan City of Rome, with a population of 4,355,725 residents, is the most populous metropolitan city in Italy. Its metropolitan area is the third-most populous within Italy. Rome is located in the central-western portion of the Italian Peninsula, within Lazio (Latium), along the shores of the Tiber. Vatican City (the smallest country in the world) is an independent country inside the city boundaries of Rome, the only existing example of a country within a city. Rome is often referred to as the City of Seven Hills due to its geographic location, and also as the "Eternal City". Rome is generally considered to be the "cradle of Western Christian culture and civilization", and the center of the Catholic Church. Rome's history spans 28 centuries. While Roman mythology dates the founding of Rome at around 753 BC, the site has been inhabited for much longer, making it a major human settlement for almost three millennia and one of the oldest continuously occupied cities in Europe. The city's early population originated from a mix of Latins, Etruscans, and Sabines. Eventually, the city successively became the capital of the Roman Kingdom, the Roman Republic and the Roman Empire, and is regarded by many as the first-ever Imperial city and metropolis. It was first called The Eternal City (Latin: Urbs Aeterna; Italian: La Città Eterna) by the Roman poet

Figure 3.10. DBpedia

patterns, which represent the core elements contributing to knowledge about specific events.



mmanuel Kant (22 April 1724 - 12 ebruary 1804) was a German philosopher. He is a central figure of modern philosophy, and set the terms by which all subsequent thinkers have had to grapple. He argued that human perception structures natural laws, and that reason is the source of morality. His thought continues to hold a major influence in ...(go to Wikipedia page)

W Biography
... Immanuel Kant was born in 1724 in Königsberg, Prussia (since 1946 the city of Kaliningrad, Kaliningrad Oblast, Russia), as the fourth of nine children (four of them reached adulthood). Baptized 'Emanuel', he changed his name to 'Immanuel'[5]...

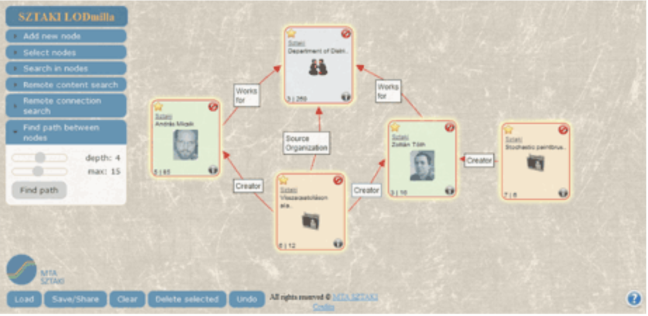
W Historical influence
... thought as different from one another as German idealism, Marxism, positivism, phenomenology, existentialism, critical theory, linguistic philosophy, structuralism, post-structuralism, and deconstructionism. [94][dubious – discuss] Historical influence[edit] Statue of Immanuel Kant in Kaliningrad (Königsberg), Russia. Replica by Harald Haacke (de) of the original by Christian Daniel Rauch lost in 1945. During

Figure 3.11. Aemoo

LODmilla [124, 125] (fig. 3.12) provides a link-node navigation where users can search and extract data associations that are hidden inside the linked data with the help of nodes. LODmilla interface does not display the data that is underlying the documents.

Tarsier [133](fig. 3.13) provides interactive 3D graph visualization of LD sources.

LODmilla (2014)



NODE-LINK VISUALIZATION KNOWLEDGE EXPLORATION INCREMENTAL VISUALIZATION DIGITAL LIBRARY

Figure 3.12. LODmilla

This tool is based on the metaphor of semantic planes. A semantic plane consists of elements that share a common concept (e.g. a semantic plane can hold all elements belonging to the same class, all the elements with the same property). The semantic planes can be created and split thanks to user interface commands. Users (leveraging the 3D environment) can personalize the layout by moving elements on different planes and get different perspectives by rotating the graph. The user interface presents several boxes and features for manipulating the graph by adding or removing elements. Boxes contain buttons for moving selected elements to different semantic planes. Resources are drawn on the graph as spheres and are placed over concentric circumferences. Generally, classes are placed over smaller circumferences, while instances are placed over wider ones. Tarsier adopts a visual schema for differentiating heterogeneous elements that are not labelled, so users must click on nodes or edges to understand what they represent. This tool was created to display graphs over different planes and accomplished that via creating a 3D-graph visualization environment. The graphs can only be created after the insertion of a SPARQL query. However, they are not expandable, and exploratory searches are impossible.

Visual query composition

In a different context, some works study the problem of node-link visual query composition. These tools allow users without prior knowledge of Semantic Web technologies to express SPARQL queries using visual graph representations easily.

SPARQLFilterFlow [141] (fig. 3.14) is a Web tool based on a filter/flow model. The tool allows users to formulate SPARQL queries using graphical elements in a

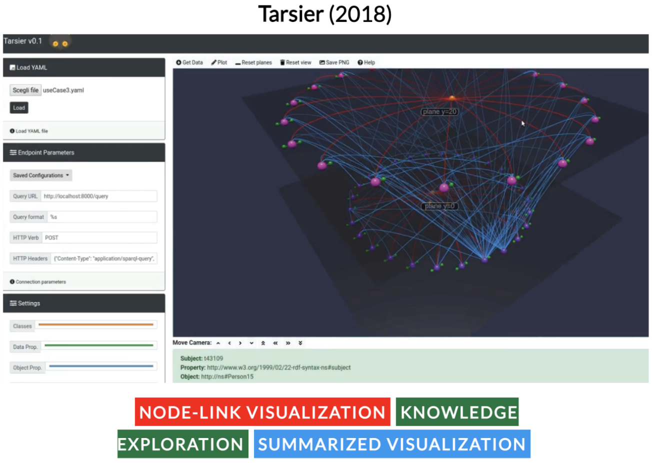


Figure 3.13. Tarsier

tree-based visualization.

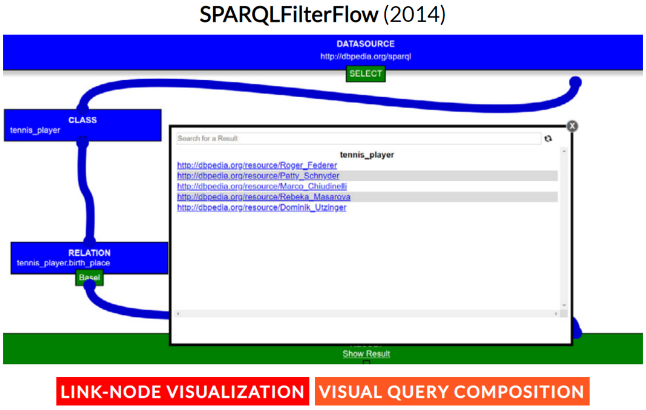

SPARQLFilterFlow (2014)
DATASOURCE
http://dbpedia.org/sparql
SELECT
CLASS
tennis_player
RELATION
tennis_player.birth_place
Basel
Search for a Result
tennis_player
http://dbpedia.org/resource/Roger_Federer
http://dbpedia.org/resource/Patty_Schnyder
http://dbpedia.org/resource/Marco_Chiudinelli
http://dbpedia.org/resource/Rebeka_Masarova
http://dbpedia.org/resource/Dominik_Utzinger
Show Result
LINK-NODE VISUALIZATION
VISUAL QUERY COMPOSITION

Figure 3.14. SPARQLFilterFlow

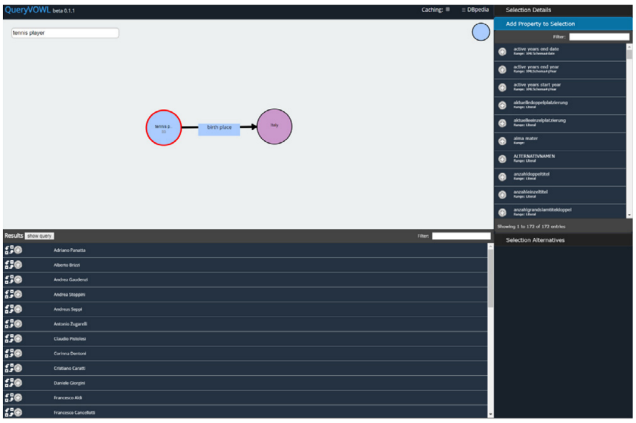
Similarly, in **QueryVOWL** [127, 128] (fig. 3.15) the user uses visual elements based on VOWL graphical elements to construct graphs that are transformed into SPARQL queries. This tool's primary focus is query construction rather than data exploration.

3.5.2 By type of information

This subsection lists tools classified by the type of information represented:

- Data Visualization. Tools made to visualize the actual data.

QueryVOWL (2015)



active years end date
active years end year
active years start year
aktuelledoppelplatzierung
aktuelleeinzelplatzierung
alma mater
ALTERNATIVNAMEN
anzahldoppeltitel
anzahleinzeltitel
anzahlgrandslamtiteldoppel

Showing 1 to 172 of 172 entries

Adriano Panatta
Alberto Brizzi
Andrea Gaudenzi
Andrea Stoppini
Andreas Seppi
Antonio Zugarelli
Claudio Pistolesi
Corinna Dentoni
Cristiano Caratti
Daniele Giorgini
Francesco Aldi
Francesco Cancellotti

LINK-NODE VISUALIZATION VISUAL QUERY COMPOSITION MODEL VISUALIZATION

Figure 3.15. QueryVOWL

- Model visualization. Tools that show data models (i.e., schemas and ontologies).
- Data to model visualization. Tools that, starting from the data, extract and visualize the underlying data model.

The tools representing each category are described below.

Data visualization

Knowledge exploration tools form a particular category of seeking information on many knowledge bases to reveal related information to the searcher and retrieve what users are looking for. With this search category, the final targets of the search are not known, and the goal itself is not defined. Therefore, a set of additional activities, for instance, learning, exploration and evaluation, are accessible through this category.

Discovery Hub [138] (fig. 3.16) is an advanced tool for knowledge presentation and exploration. Although similar to QueryVOWL, Discovery Hub is limited to the knowledge from DBpedia.

Model visualization

Ontodia [12] (fig. 3.17) is a Web-based ontology, and semantic data set visualization tool with additional functionality in sharing and distributing resulting diagrams. Ontodia utilizes the 2D node-link visualization approach adopting a UML-inspired way of displaying additional information about the node. In terms of

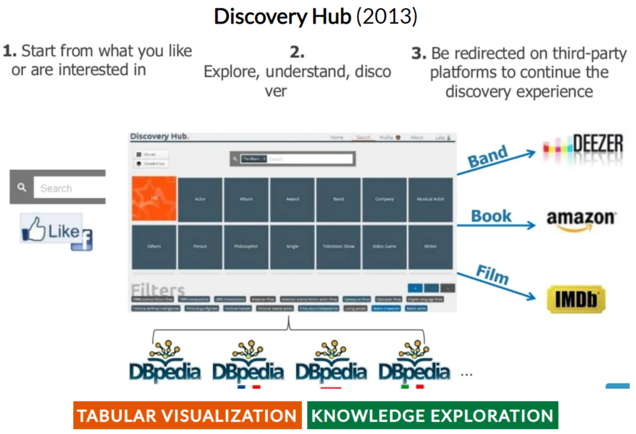

Discovery Hub (2013)
1. Start from what you like or are interested in
2. Explore, understand, disco ver
3. Be redirected on third-party platforms to continue the discovery experience
Search
Like
Band
Book
Film
DEEZER
amazon
IMDb
Filters
DBpedia
DBpedia
DBpedia
DBpedia
...
TABULAR VISUALIZATION
KNOWLEDGE EXPLORATION

Figure 3.16. Discovery Hub

layouts, the tool offers force-directed and grid layouts. It includes the hierarchical relationships view that displays the parent-child relationships between classes in a tree layout. Since the tool claims the ability to visualize semantic datasets, it also enables drag-n-dropping instances on the diagram. The view of the diagram could be freely adjusted by a user through drag-n-dropping additional items on the canvas, rearranging them, removing nodes from the graph, and turning links between nodes on and off. Ontodia supports data exploration capability so the user can sort out the nodes related to the selected node. From the instance panel, he can drag-n-drop one or several related nodes on canvas, thus expanding the graph and exploring the ontology. The tool has unique diagram management features that allow users to publish the fixed URL of the diagram on the Web, share it with others via the email address or lock it to themselves. The tool introduces the data source entity, and access to later can be managed similar to controlling access to a diagram. Searching and filtering are fully available for classes, instances and links. Ontodia was designed to simplify the visualization of ontologies and semantic data. For this reason, some of the OWL constructs were omitted, and only the basic ones are kept on the graph.

TopBraid Composer[11] (fig. 3.18), is designed mainly for ontology editing. Its visualization, a side feature, is not particularly convenient for simple tasks for several reasons. Composer's full version is paid, while Ontodia is a free service, which could be a decisive factor for researchers. Moreover, TopBraid exists in the Eclipse environment, which is not comfortable for non-programmers, while Ontodia is deliberately designed for non-programming users. Ontologies can be displayed in the TopBraid Composer using a UML-inspired graph visualization method with a

[11] `https://www.w3.org/wiki/TopBraid`

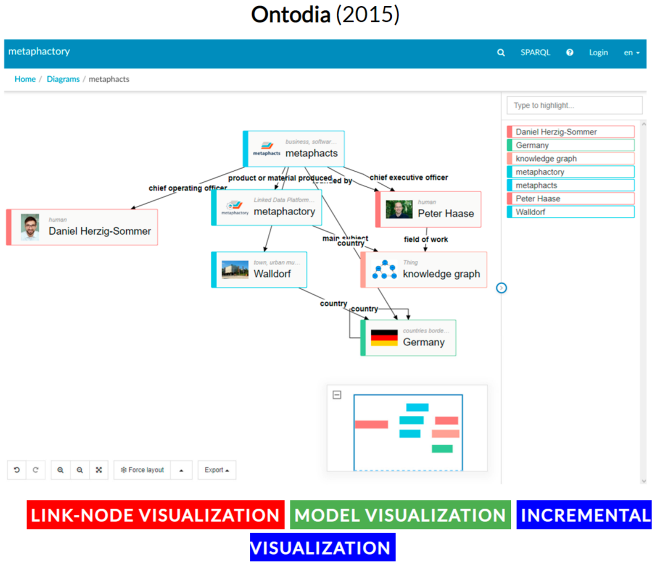


Figure 3.17. Ontodia

horizontal or vertical tree layout backed up by a classic indented list view. Classes and properties are depicted by nodes connected with oriented edges labelled with the predicate name they represent. There is also a possibility of displaying class data and annotation properties with their values inside the class node. The graph visualization shows all types of entities as nodes and predicates as edges connecting them. Datatype property values can be displayed inside the node of the class they are describing. Restriction and complex anonymous classes are depicted with analogous mathematical symbols. The visualization is done rather at the RDF level, similarly to Graffoo or NavigOWL, so even owl:Class is shown as a separate node and every class is connected to it through an rdf:type edge.

WebVOWL [129] (fig. 3.19) is a Web application for the user-oriented visualization of ontologies. It implements the Visual Notation for OWL Ontologies (VOWL) by providing graphical depictions for elements of OWL that are combined to a force-directed graph layout representing the ontology (Lohmann et al., Reference Lohmann, Negru, Haag and Ertl2016). Interaction techniques allow users to explore the ontology and customize the visualization. The VOWL visualizations are automatically generated from JSON files into which the ontologies need to be converted. A Java-based OWL2VOWL converter is provided along with WebVOWL. The force-directed graph layout uses a physics simulation where the forces are iteratively applied, resulting in an animation that dynamically positions the graph nodes. The energy of the forces cools down in each iteration, and the layout animation stops automatically after some time to provide a stable graph visualization. WebVOWL

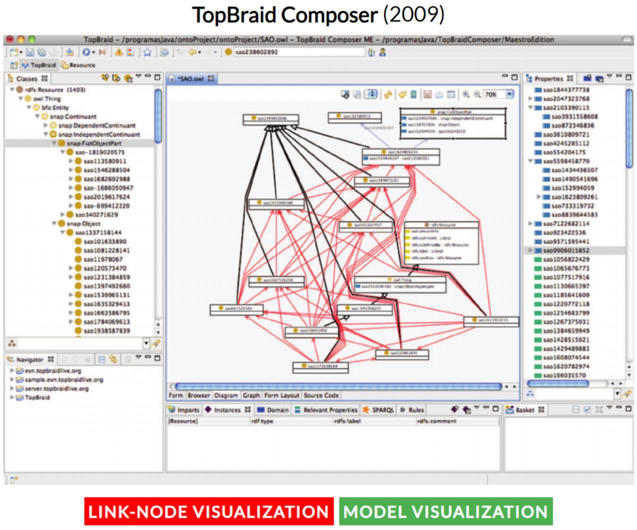


Figure 3.18. TopBraid Composer

renders the graphical elements according to the VOWL specification.

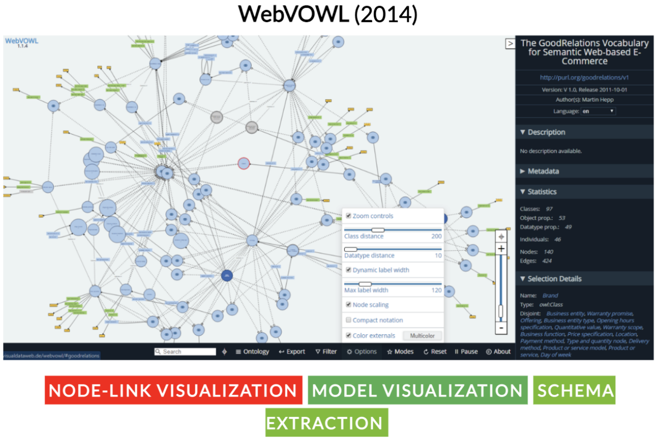


Figure 3.19. WebVOWL

Data to model visualization (schema extraction)

Recently, the problem of data to model visualization, namely the schema extraction, has been examined in the context of LD. **VizLOD** [139] (fig. 3.20), **LD-VOWL** [122] (fig. 3.21), and **RDF2Graph** [140] use SPARQL queries to process RDF triples to infer schema information. The tools process the LD to infer the ontology schema. The tools first identify and present the most representative concepts, using several methods and assumptions, e.g., consider the classes with a more significant number of instances as representative. Then, the ontology schema is (progressively) visualized as a graph, offering several interactive operations.

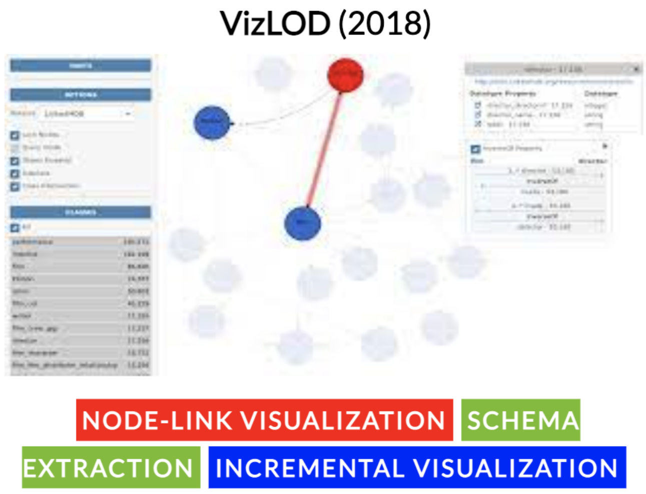
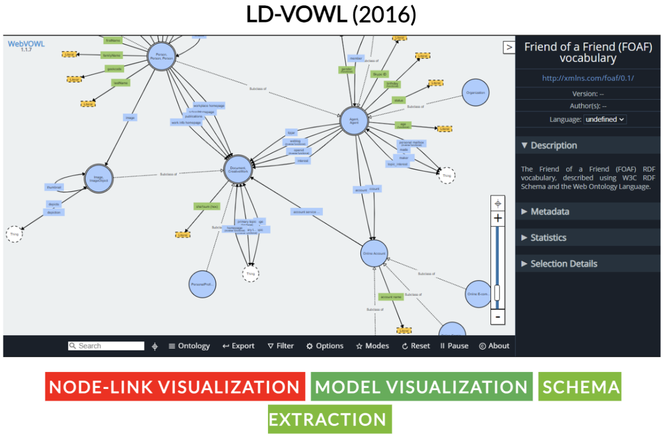


Figure 3.20. VizLOD



Friend of a Friend (FOAF) vocabulary

http://xmlns.com/foaf/0.1/

Version: –

Author(s): –

Language: undefined

The Friend of a Friend (FOAF) RDF vocabulary, described using W3C RDF Schema and the Web Ontology Language.

Figure 3.21. LD-VOWL

3.5.3 Complexity reduction strategies

Viewing a large number of data objects is a challenging task. Even in small datasets, providing an overview of the dataset can be extremely difficult; in both cases, information overload is a common problem. Consequently, visual scalability is a fundamental requirement of modern systems, which must effectively support data reduction/abstraction on many data objects.

Two strategies have been identified for reducing the number of information displayed:

- **Navigational visualization**. The visualization in many user interfaces is focused around a specific data object, typically a resource. The user can see the "neighborhood" of the current resource and can navigate to directly related resources. This strategy is often used in the tabular interaction paradigm.

- **Incremental visualization**. The paradigm of incremental visualization is often adopted in dynamic node-link user interfaces. The user controls a workspace in which them can add or remove views of specific data objects from the data set as needed. There are often shortcuts to visualize data objects related to the objects already in view.

- **Summarized Visualization**. In order to offer an overview of a dataset, while avoiding the problem of overplotting (related to visual information overloading) in large graph visualizations, several tools use data reduction techniques to provide graph summaries.

The tools representing each category are described below.

Navigational Visualization

Far from the paradigm of simple linear search results lists, new and more expressive navigation features, such as node-link views, cluster maps, geographic maps and timelines, support the user in the perception of information. Due to the high diversity of relationships between entities, interfaces must be highly generic. This requirement needs methods to structure information visually based on the user's interests. Therefore, it is necessary to assign specific relevance to related entities.

For example, there are more than 600 facts (RDF triples) to view information about the DBpedia entity "Imperatore Augusto". This amount of information cannot be presented to the user at a glance. Furthermore, each user may have different preferences. Heuristics based on the statistical and semantic analysis of the underlying RDF graph structure are applied to classify related entities based on their relevance. Therefore, relevance rankings need to be customized. User behavior can be monitored by analyzing the log file. With the user's preferences, it is possible to generate a profile and map it to a LOD sub-chart, representing the user's interests. This allows:

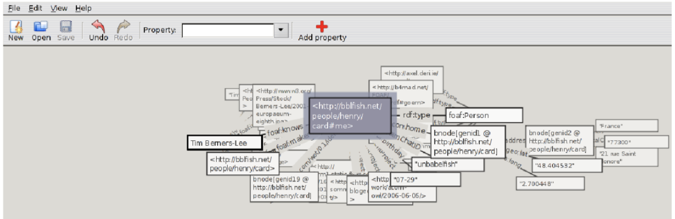
- subjective relevance ranking

- personalized search recommendations.

Incremental visualization

Fenfire [137] (fig. 3.22), **LodLive** [123] (fig. 3.23), and **LodView**[12] are Web exploratory tools that allow users to browse LD using interactive graphs. The user can explore LD by following the links starting from a given URI or a SPARQL endpoint.

Fenfire (2008)


File
Edit
View
Help
New
Open
Save
Undo
Redo
Property:
Add property
<http://bblfish.net/people/henry/card#me>
rdf:type
foaf:Person
foaf:knows
foaf:maker
Tim Berners-Lee
<http://bblfish.net/people/henry/card>
<http://www.w3.org/Press/Stock/Berners-Lee/2001-europaeum-eighth.jpg>
<http://axel.deri.ie/
< http://b4mad.net/FOAF/rdf#goern>
bnode[genid1 @ http://bblfish.net/people/henry/card]
bnode[genid2 @ http://bblfish.net/people/henry/card]
bnode[genid19 @ http://bblfish.net/people/henry/card]
"unbabelfish"
"07-29"
"France"
"77300"
"21 rue Saint onore"
"48.404532"
"2.700448"
geo:lat
geo:long
owl/2006-06-05/>

NODE-LINK VISUALIZATION KNOWLEDGE EXPLORATION INCREMENTAL VISUALIZATION

Figure 3.22. Fenfire

LodLive is a linked data visualization tool that allows the user to explore a dataset starting from a single resource and then to widen the research by expanding the properties of the elements displayed on the screen. LodLive can retrieve information from several endpoints after providing the URI of a resource. It also provides an autocompletion feature to the users for choosing the URI of the resource of interest. Resources are drawn on the screen through circles that contain the value of the property rdfs:label; then, other smaller circles are drawn concentrically. Every circle represents an object connected to the central resource. If two objects are connected through different properties, the link connecting them will report all labels related to those properties. Smaller circles can be further expanded too. Adding new resources via keyword or URI is impossible without deleting the complete graph. LodLive emphasizes inverse properties, owl:sameAs relations, images, and geographic data. For each of them, LodLive provides the special treatment. Inverse properties and owl:sameAs are represented through particular smaller circles, images are collected and shown in a reserved panel while geographic data are geolocalized on a map. A helpful feature is to collect information from different endpoints about resources referring to the same thing (connected through owl:sameAs property).

Summarized Visualization

H-BOLD [121] (fig. 3.24), the acronym for High-level visualization over Big Linked Open Data, is the successor of LODeX. It aims to help the general user,

[12] `https://lodview.it`

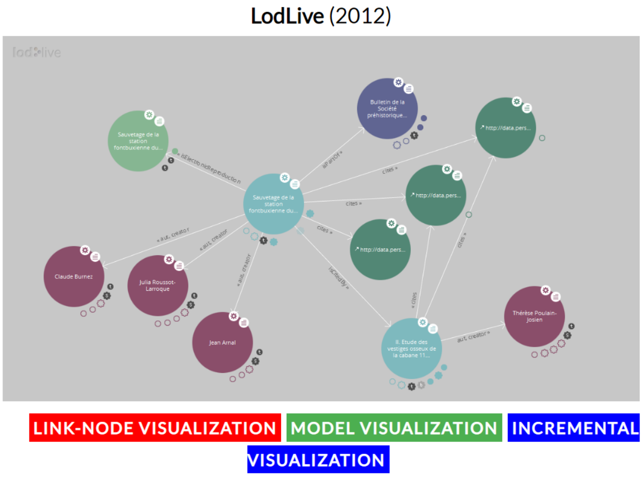
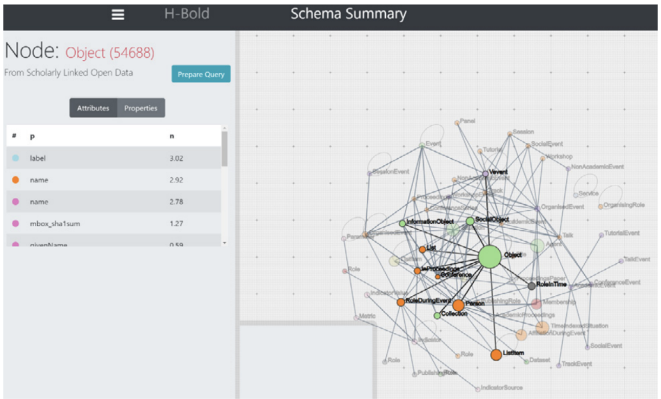


Figure 3.23. LodLive

without any knowledge about SPARQL and the content of a dataset, to explore significant sources of linked open data. It offers incremental multilevel exploration, using a community detection algorithm to construct the abstract levels effectively. The visual interpretability of a dataset strongly depends on the amount of information shown on the screen. For this reason, H-BOLD proposes an abstract visualization of the dataset where classes have been aggregated into clusters. Users can further incrementally explore classes. For summarized representation and exploration, H-BOLD collects information (number of triples, class list, property list, and class relations) about datasets in a data store and uses that information for generating a “Schema Summary”. If the number of classes is high, a community detection algorithm is executed to shrink the graph into the “Cluster Schema”. Cluster and Schema summaries can be expanded on-the-fly for detailed visualization of the dataset. The tool can create a fast and compact overview of the content of a SPARQL endpoint.

In the same context, **RDF4U** [130] (fig. 3.25) offers graph visualization over summarized graphs. Researchers noticed that graphs are potent instruments for enhancing human comprehension due to their readability and synthesis, but this is true when graphs display only relevant information. The RDF4U visualization approach is based on combining graph simplification, triple ranking, and property selection. The tool automatically analyzes data collected after querying and searches for redundant information during graph simplification. Generally, redundant information came from equivalent (e.g. owl:sameAs), transitive (e.g. skos:broaderTransitive) and hierarchical classification (e.g. rdf:type, owl:subClassOf) relations. Some endpoints are equipped with reasoning functionality and return inferred triples. One common action performed in this step is merging owl:sameAs nodes in a unique node. The

H-BOLD (2018)



Node: Object (54688)

From Scholarly Linked Open Data

#	p	n
●	label	3.02
●	name	2.92
●	name	2.78
●	mbox_sha1sum	1.27
●	givenName	0.59

NODE-LINK VISUALIZATION KNOWLEDGE EXPLORATION INCREMENTAL VISUALIZATION SUMMARIZED VISUALIZATION

Figure 3.24. H-BOLD

tool aimed to show information at different times, common information at first and then show topic-specific information as the user wishes. The initial graph contains only "Common Information" properties, but "Topic-Specification" ones can be added to the graph through a simple button or by adjusting a range bar above the graph. The user interface is minimal, helping users to focus on the graph. It is possible to start research both by the URI of a resource or by a SPARQL query over a selected SPARQL endpoint. This tool resulted in being adaptive for different kinds of users. Both non-technical users and domain experts can visualize the information they need. General users can hide domain-specific information by focusing on general ones, while domain experts can directly visualize more specific information by hiding the general ones.

The **Graphless** [118] (fig. 3.26) visualization tool generates summaries based on statistical data, e.g., nodes' connectivity degree and property frequency.

3.5.4 Advanced functionalities

For advanced functionalities, we want to indicate specific features that go beyond those commonly offered by a specific interaction paradigm (such as view manipulation, display properties, etc.).

Paths discovery

Some tools focus on analyzing the associations between LD entities. In this context, associations (i.e., relationships) are considered the paths connecting the

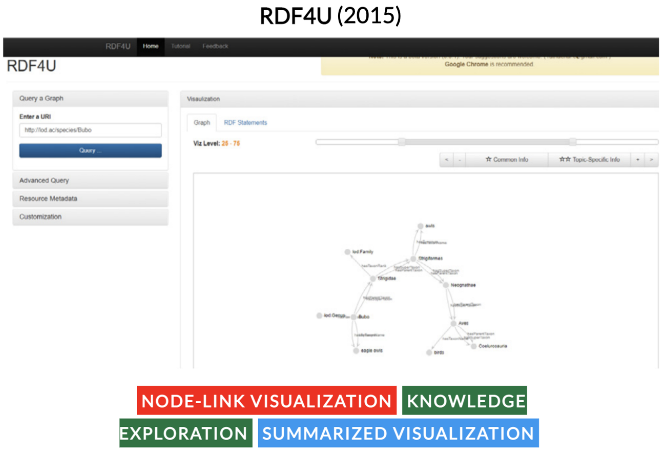


Figure 3.25. RDF4U

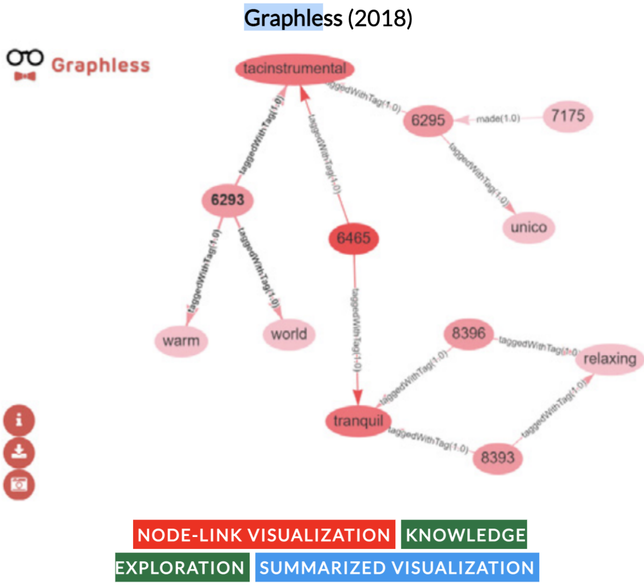


Figure 3.26. Graphless

entities in the LD graph.

RelFinder [132] (fig. 3.27) is a Web-based tool that offers interactive discovery and visualization of relationships between selected LD resources. The main idea behind RelFinder is that even domain experts do not know everything about their research fields. RelFinder is a way to avoid missing essential relations in the Semantic Web. RelFinder implements the incremental reduction strategy: filters can be applied to increase or reduce the number of relationships that are shown in the graph and to focus on certain aspects of interest

Another critical aspect regards the visualization of extracted relations. The higher the number of relations between different objects is, the easier it is to overlook some of them. RelFinder includes four sequential steps:

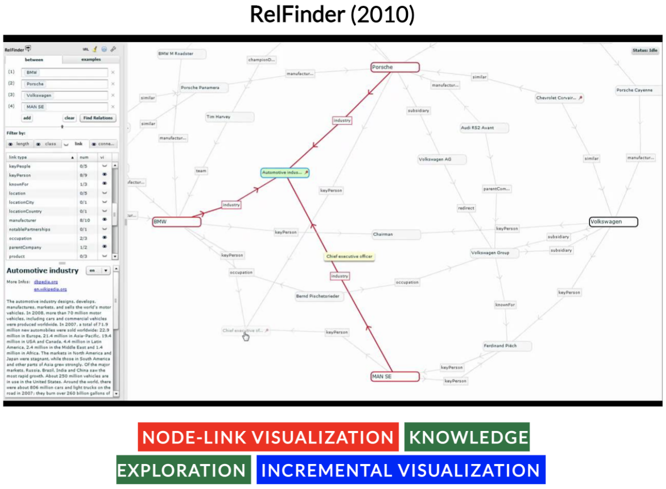
- Object Mapping: in this step, requested resources must be mapped in semantic elements. Where this is not possible at first, it must be considered to help users via an auto-completion feature so that they can manually disambiguate the elements he is looking for.

- Relationship Search: after the resources have been selected, the relationship search starts over the selected dataset. Datasets are crawled to find the highest number of relations since each could be important in certain situations.

- Visualization: in this step, relationships are presented to the user. The spatial limitation is one of the most significant issues in enhancing human cognition. The visualization layout should help the user understand what has been extracted in the previous step. Valuable features that help human cognition could include statistical information about the extracted relationship and some aggregated visualization.

- Interactive Exploration: Highlighting interesting relations or interactive visualization like fish-eye zooming could boost understanding of relationships.

RelFinder interface presents a window where the user can select the resources of interest; he can also select more than two resources. The tool allows querying different endpoints by adding new URLs in the setting menu.



(1) BMW
(2) Porsche
(3) Volkswagen
(4) MAN SE

link type	num	vi
keyPeople	0/5	
keyPerson	8/9	
knownFor	1/3	
location	0/5	
locationCity	0/1	
locationCountry	0/1	
manufacturer	8/10	
notablePartnerships	0/1	
occupation	2/3	
parentCompany	1/2	
product	0/3	

Automotive industry en

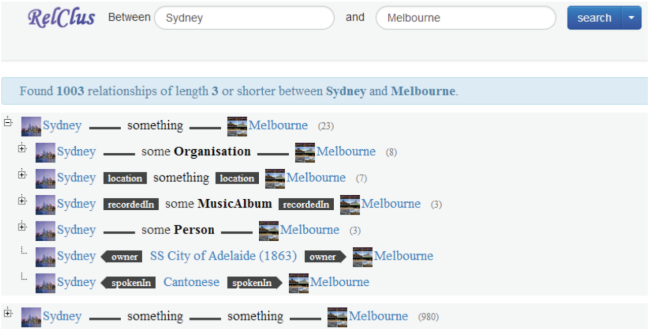
More Infos: dbpedia.org
en.wikipedia.org

The automotive industry designs, develops, manufactures, markets, and sells the world's motor vehicles. In 2008, more than 70 million motor vehicles, including cars and commercial vehicles were produced worldwide. In 2007, a total of 71.9 million new automobiles were sold worldwide: 22.9 million in Europe, 21.4 million in Asia-Pacific, 19.4 million in USA and Canada, 4.4 million in Latin America, 2.4 million in the Middle East and 1.4 million in Africa. The markets in North America and Japan were stagnant, while those in South America and other parts of Asia grew strongly. Of the major markets, Russia, Brazil, India and China saw the most rapid growth. About 250 million vehicles are in use in the United States. Around the world, there were about 806 million cars and light trucks on the road in 2007; they burn over 260 billion gallons of

Figure 3.27. RelFinder

RelClus [135] (fig. 3.28) uses class/property hierarchies to generate hierarchies of paths.

RelClus (2013)



RelClus Between Sydney and Melbourne search

Found **1003** relationships of length **3** or shorter between **Sydney** and **Melbourne**.

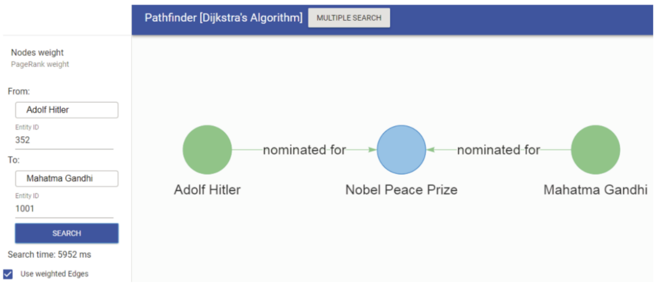
- Sydney —— something —— Melbourne (23)
  - Sydney —— some **Organisation** —— Melbourne (8)
  - Sydney location something location Melbourne (7)
  - Sydney recordedIn some **MusicAlbum** recordedIn Melbourne (3)
  - Sydney —— some **Person** —— Melbourne (3)
  - Sydney owner SS City of Adelaide (1863) owner Melbourne
  - Sydney spokenIn Cantonese spokenIn Melbourne
- Sydney —— something —— something —— Melbourne (980)

TABULAR VISUALIZATION KNOWLEDGE EXPLORATION NAVIGATIONAL VISUALIZATION

Figure 3.28. RelClus

WiSP [134] (Weighted Shortest Paths for RDF Graphs) (fig. 3.29) uses weighted shortest path algorithms to identify and present the most relevant paths between two resources. The weights are computed based on several metrics, e.g., PageRank and node degree. The researchers' main goal behind WISP is to investigate weighting schemas for RDF graphs such that the weighted shortest paths are more interesting than simple shortest paths. To help users select the resources of interest, WiSP provides an autocompletion feature. The web application allows users to find relations only in the Wikidata triplestore.

WiSP (2018)



Pathfinder [Dijkstra's Algorithm] MULTIPLE SEARCH

Nodes weight
PageRank weight

From:
Adolf Hitler
Entity ID
352

To:
Mahatma Gandhi
Entity ID
1001

SEARCH

Search time: 5952 ms

☑ Use weighted Edges

NODE-LINK VISUALIZATION VISUAL QUERY COMPOSITION SUMMARIZED VISUALIZATION

Figure 3.29. WiSP

The **LODVader** [136] (LOD Visualization, Analytics and DiscovERy) system visualizes linked datasets, following the graph layout used in the LOD cloud visualization, where the nodes represent the datasets and the edges the links between them. The system is built on top with an index structure and a database system, which allow efficient link analysis and extraction techniques between the different datasets and similarity computations.

3.6 Semantic annotations

Semantic annotations tools aim to annotate documents with entities, classes, topics or facts, typically based on an existing ontology/KB. Some works on Semantic Annotation include extraction and linking of entities and/or concepts (though not typically relations).

Methods to semantically enrich unstructured content can be classified in three main approaches [93] :

- *manual*;
- *automatic*;
- *semi-automatic*.

3.6.1 Manual

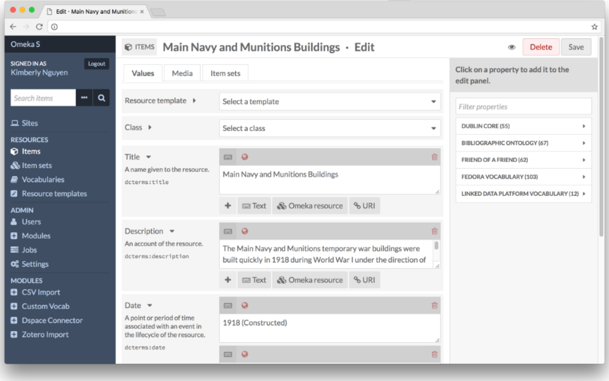
In the *manual* approach humans are fully in charge with the semantic annotation of content.

Omeka S[13] (fig. 3.30), for example, is a platform to connect digital cultural heritage collections with other online resources, and **SenTag** [96], which enables tagging a corpus of documents through an intuitive and easy-to-use user interface. In manual systems, quality errors are reduced to human input errors, but the throughput of the process is limited by the amount time knowledgeable people can contribute to the process. Furthermore, the possible divergence in the criteria used for classification by different users' needs also be taken into account.

AKTiveMedia [142] (fig. 3.31) is a user-centric system for annotating documents with the support of text, images and HTML documents (containing both text and images) with ontology-based and free-text annotations. Both author and reader can perform annotations allowing the utilization of different ontologies. The annotations are not stored in the document but separately with authorship allowing users to share comments and annotations with other community members using a centralized server. Most annotations are done manually, but various techniques are available to reduce the effort of annotating.

[13] `https://omeka.org`

Omeka S (2022)



Main Navy and Munitions Buildings · Edit

Click on a property to add it to the edit panel.

Resource template: Select a template

Class: Select a class

Title — A name given to the resource. `dcterms:title`: Main Navy and Munitions Buildings

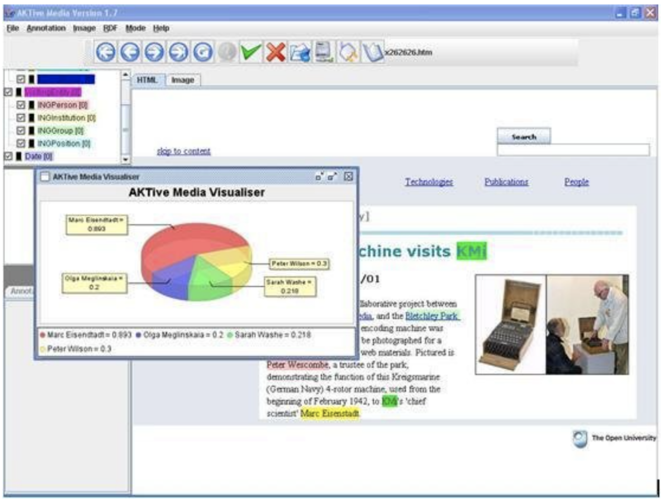
Description — An account of the resource. `dcterms:description`: The Main Navy and Munitions temporary war buildings were built quickly in 1918 during World War I under the direction of

Date — A point or period of time associated with an event in the lifecycle of the resource. `dcterms:date`: 1918 (Constructed)

NAMED ENTITY RECOGNITION SEMANTIC ENRICHMENT MANUAL SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.30. Omeka S

AKTiveMedia (2006)



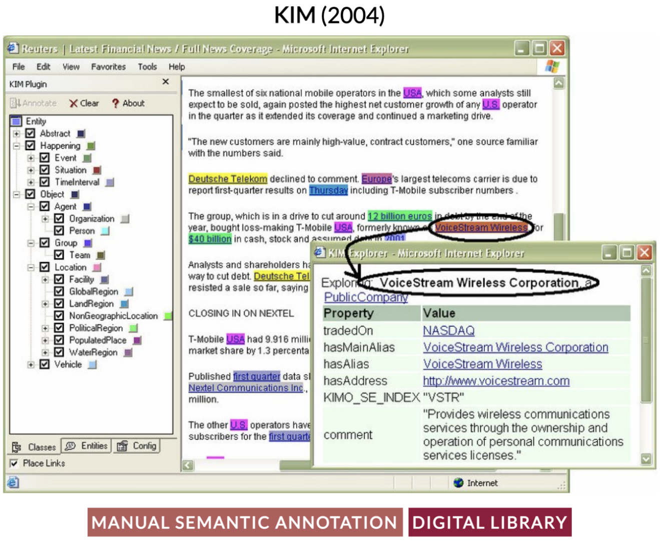
● Marc Eisendtadt = 0.893 ● Olga Meglinskaia = 0.2 ● Sarah Washe = 0.218 ● Peter Wilson = 0.3

...llaborative project between ...dia, and the Bletchley Park ... encoding machine was ... be photographed for a ... web materials. Pictured is Peter Wescombe, a trustee of the park, demonstrating the function of this Kriegsmarine (German Navy) 4-rotor machine, used from the beginning of February 1942, to KMi's 'chief scientist' Marc Eisenstadt.

MANUAL SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.31. AKTiveMedia

KIM [115] (Knowledge and Information Management) (fig. 3.32) platform contains an ontology, a knowledge base, an automatic Semantic Annotation, indexing, and a retrieval server. Like SemTag, KIM focuses on assigning to the entities in the corpus links to their semantic descriptions, provided by the KIMO ontology that, apart from containing named entity classes and their properties, is pre-populated with a large number of instances. Created annotations are linked to the entity type and the exact individual in the knowledge base. KIM offers an infrastructure capable of scalable and customizable information extraction, annotation, and document management, based on GATE (the General Architecture for Text Engineering). In order to provide a basic level of performance and allow easy bootstrapping of applications, KIM is equipped with an upper-level ontology and a knowledge base providing extensive coverage of entities of general importance. From a technical point of view, the platform allows KIM-based applications to use it for automatic Semantic Annotation, content retrieval based on semantic restrictions, and querying and modifying the underlying ontologies and knowledge bases.



KIM (2004)

The smallest of six national mobile operators in the USA, which some analysts still expect to be sold, again posted the highest net customer growth of any U.S. operator in the quarter as it extended its coverage and continued a marketing drive.

"The new customers are mainly high-value, contract customers," one source familiar with the numbers said.

Deutsche Telekom declined to comment. Europe's largest telecoms carrier is due to report first-quarter results on Thursday including T-Mobile subscriber numbers .

The group, which is in a drive to cut around 12 billion euros in debt by the end of the year, bought loss-making T-Mobile USA, formerly known as VoiceStream Wireless, for $40 billion in cash, stock and assumed debt in 2001.

Analysts and shareholders ha
way to cut debt. Deutsche Tel
resisted a sale so far, saying

CLOSING IN ON NEXTEL

T-Mobile USA had 9.916 milli
market share by 1.3 percenta

Published first quarter data s
Nextel Communications Inc.,
million.

The other U.S. operators have
subscribers for the first quarte

Exploring: VoiceStream Wireless Corporation, a PublicCompany

Property	Value
tradedOn	NASDAQ
hasMainAlias	VoiceStream Wireless Corporation
hasAlias	VoiceStream Wireless
hasAddress	http://www.voicestream.com
KIMO_SE_INDEX	"VSTR"
comment	"Provides wireless communications services through the ownership and operation of personal communications services licenses."

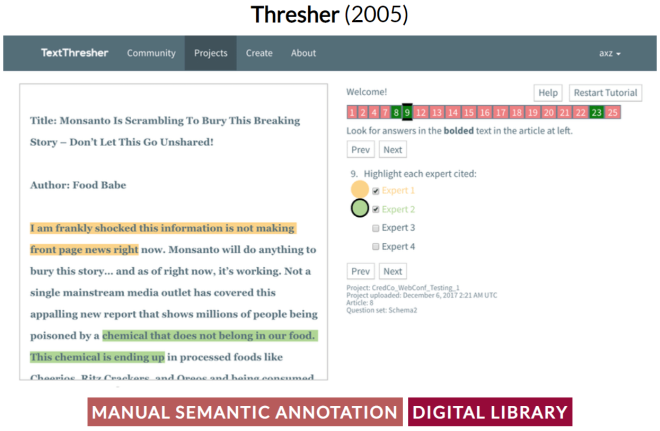
MANUAL SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.32. KIM

Thresher [143] (fig. 3.33) allows end-users, instead of content providers, to unwrap the semantic structures nested inside Web pages. Thresher presents a web interface allowing non-technical users to mark up examples of a particular class quickly. By analogy, thresher learns from these examples, so it can induce wrappers automatically that can be applied to the same page or "similar" web pages. Thresher is aimed at Web pages with similar content (same type of object). Usually, web pages are fed relational data through a template and, by analogy, extract the corresponding information.

3.6.2 Automatic

Automatic methods use a variety of techniques (including machine learning algorithms and natural language processing) in order to allow a machine to derive semantic information from the unstructured content, with minimal user intervention.



Title: Monsanto Is Scrambling To Bury This Breaking Story – Don't Let This Go Unshared!

Author: Food Babe

I am frankly shocked this information is not making front page news right now. Monsanto will do anything to bury this story... and as of right now, it's working. Not a single mainstream media outlet has covered this appalling new report that shows millions of people being poisoned by a chemical that does not belong in our food. This chemical is ending up in processed foods like Cheerios, Ritz Crackers, and Oreos and being consumed

Look for answers in the **bolded** text in the article at left.

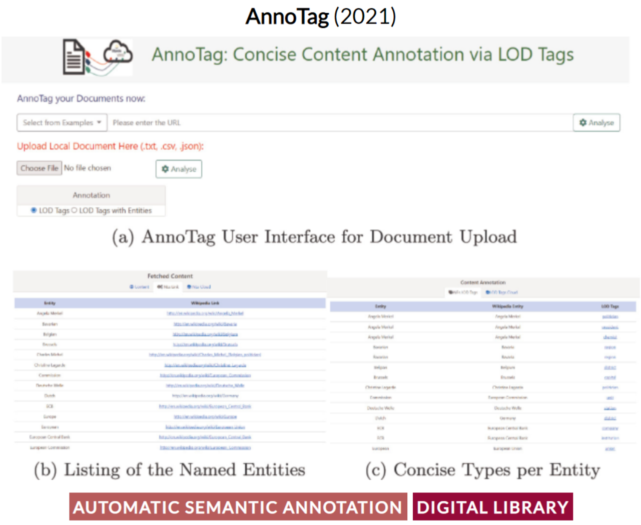
9. Highlight each expert cited:

☑ Expert 1
☑ Expert 2
□ Expert 3
□ Expert 4

Project: CredCo_WebConf_Testing_1
Project uploaded: December 6, 2017 2:21 AM UTC
Article: 8
Question set: Schema2

Figure 3.33. Thresher

AnnoTag [94] (fig. 3.34), for example, provides concise content annotations by employing entity-level analytics in order to derive semantic descriptions in the form of tags. The Arca [76] (fig. 4.1) system, thanks to the automatic association of unstructured content with concepts in a knowledge graph (KG), allow complex queries on data and visualization of semantic associations connecting concepts and documents.



AnnoTag your Documents now:

Upload Local Document Here (.txt, .csv, .json):

No file chosen

● LOD Tags ○ LOD Tags with Entities

(a) AnnoTag User Interface for Document Upload

(b) Listing of the Named Entities

(c) Concise Types per Entity

Figure 3.34. AnnoTag

OpenCalais[14] (fig. 3.6) is a web service which automatically creates rich

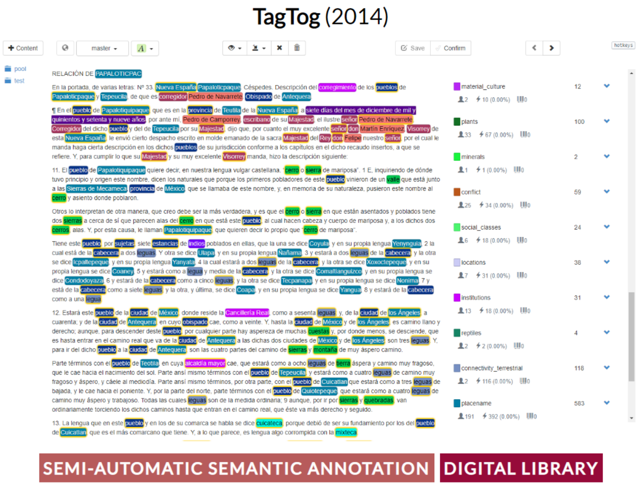
[14] http://viewer.opencalais.com

semantic metadata from an unstructured text source. OpenCalais performs natural language processing and uses machine-learning techniques to define entities in text. Entities are divided into named entities, facts, and events. Constructing maps (or graphs or networks) linking documents to people, companies, places, and other entities is possible. OpenCalais is offered for free but with a daily limit of requests.

3.6.3 Semi-automatic

Semi-automatic approaches as the one presented here aim to get the best of both worlds, by providing some form of collaboration between the machine, providing automatic annotation, and the human experts, who are still able to control the final result of the process.

An example is the one of **tagtog**[15] [95](fig. 3.35), which is a collaborative tool to annotate texts. It allows searching for documents and entities, but lacks the capability of visualising relationships between entities.


TagTog (2014)

RELACIÓN DE PAPALOTICPAC

En la portada, de varias letras: Nº 33. Nueva España Papaloticpaque. Céspedes. Descripción del corregimiento de los pueblos de Papaloticpaque y Tepeucila, de que es corregidor Pedro de Navarrete. Obispado de Antequera.

¶ En el pueblo de Papalotiquipaque, que es en la provincia de Teutila de la Nueva España, a siete días del mes de diciembre de mil y quinientos y setenta y nueve años, por ante mí, Pedro de Camporrey, escribano de su Majestad, el ilustre señor Pedro de Navarrete, Corregidor del dicho pueblo y del de Tepeucila por su Majestad, dijo que, por cuanto el muy excelente señor don Martín Enríquez, Visorrey de esta Nueva España, le envió cierto despacho escrito en molde emanado de la sacra Majestad del Rey don Felipe nuestro señor, por el cual le manda haga cierta descripción en los dichos pueblos de su jurisdicción conforme a los capítulos en el dicho recaudo insertos, a que se refiere. Y, para cumplir lo que su Majestad y su muy excelente Visorrey manda, hizo la descripción siguiente:

11. El pueblo de Papalotiquipaque quiere decir, en nuestra lengua vulgar castellana, "cerro o sierra de mariposa". 1 E, inquiriendo de dónde tuvo principio y origen este nombre, dicen los naturales que porque los primeros pobladores de este pueblo vinieron de un valle que está junto a las Sierras de Mecameca, provincia de México, que se llamaba de este nombre, y, en memoria de su naturaleza, pusieron este nombre al cerro y asiento donde poblaron.

Otros lo interpretan de otra manera, que creo debe ser la más verdadera, y es que el cerro o sierra en que están asentados y poblados tiene dos sierras a cerca de sí que parecen alas del cerro en que está este pueblo, al cual hacen cabeza y cuerpo de mariposa y, a los dichos dos cerros, alas. Y, por esta causa, le llaman Papalotiquipaque, que quieren decir lo propio que "cerro de mariposa".

Tiene este pueblo, por sujetas, siete estancias de indios poblados en ellas, que la una se dice Coyula, y en su propia lengua Yenynguia, 2 la cual está de la cabecera a dos leguas. Y otra se dice Ulapa, y en su propia lengua Ñañama, 3 y estará a dos leguas de la cabecera; y la otra se dice Icpaltepeque, y en su propia lengua Yanyata, 4 la cual estará a dos leguas de la cabecera; y la otra se dice Xoxoctepeque, y en su propia lengua se dice Coaney, 5 y estará como a legua y media de la cabecera; y la otra se dice Comaltianguizco, y en su propia lengua se dice Condodoyaza, 6 y estará de la cabecera como a cinco leguas; y la otra se dice Tecpanapa, y en su propia lengua se dice Nonima, 7 y está de la cabecera como a siete leguas; y la otra, y última, se dice Coapa, y en su propia lengua se dice Yangua, 8 y estará de la cabecera como a una legua.

12. Estará este pueblo de la ciudad de México, donde reside la Cancillería Real, como a sesenta leguas; y, de la ciudad de los Ángeles, a cuarenta; y de la ciudad de Antequera, en cuyo obispado cae, como a veinte. Y, hasta la ciudad de México y de los Ángeles, es camino llano y derecho; aunque, para descender deste pueblo, por cualquier parte hay aspereza de muchas cuestas y, por donde menos, se desciende, que es hasta entrar en el camino real que va de la ciudad de Antequera a las dichas dos ciudades de México y de los Ángeles: son tres leguas. Y, para ir del dicho pueblo a la ciudad de Antequera, son las cuatro partes del camino de sierras y montaña de muy áspero camino.

Parte términos con el pueblo de Teotila, en cuya alcaldía mayor cae, que estará como a ocho leguas de tierra áspera y camino muy fragoso, que le cae hacia el nacimiento del sol. Parte ansí mismo términos con el pueblo de Tepeucila, y estará como a cuatro leguas de camino muy fragoso y áspero, y cáele al mediodía. Parte ansí mismo términos, por otra parte, con el pueblo de Cuicatlan que estará como a tres leguas de bajada, y le cae hacia el poniente. Y, por la parte del norte, parte términos con el pueblo de Quiotepeque, que estará como a cuatro leguas de camino muy áspero y trabajoso. Todas las cuales leguas son de la medida ordinaria; 9 aunque, por ir por sierras y quebradas, van ordinariamente torciendo los dichos caminos hasta que entran en el camino real, que éste va más derecho y seguido.

13. La lengua que en este pueblo y en los de su comarca se habla se dice cuicateca, porque debió de ser su fundamiento por los del pueblo de Cuicatlan, que es el más comarcano que tiene. Y, a lo que parece, es lengua algo corrompida con la mixteca.

material_culture	12	2 · 10 (0.00%) · 0
plants	100	33 · 67 (0.00%) · 0
minerals	2	1 · 1 (0.00%) · 0
conflict	59	25 · 34 (0.00%) · 0
social_classes	24	6 · 18 (0.00%) · 0
locations	38	7 · 31 (0.00%) · 0
institutions	31	13 · 18 (0.00%) · 0
reptiles	4	2 · 2 (0.00%) · 0
connectivity_terrestrial	118	2 · 116 (0.00%) · 0
placename	583	191 · 392 (0.00%) · 0

SEMI-AUTOMATIC SEMANTIC ANNOTATION DIGITAL LIBRARY

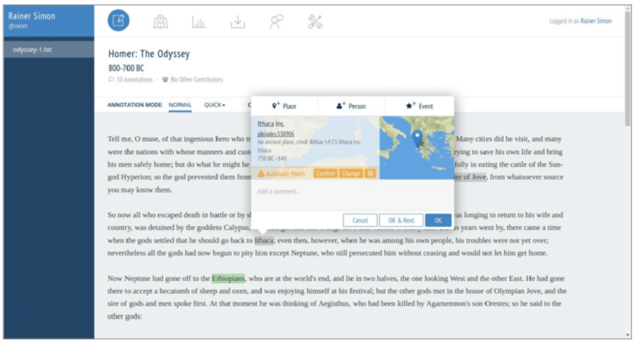
Figure 3.35. tagtog

Recogito[16] is an open source annotation tool [112] (fig. 3.36) with the aim to facilitate linkages between online resources documenting the past. It allows users to annotate text and images (i.e., ancient maps, images in digital books).

GoNTogle [144] (fig. 3.37) offers a framework for Semantic ontology-based Annotations. It is possible to annotate different document formats and allows the annotation of whole documents or fragments. Annotations are stored in a centralized ontology server, maintaining them separate from the document. This framework

[15] `https://tagtog.net`
[16] `https://recogito.pelagios.org`

Recogito (2017)



Homer: The Odyssey
800-700 BC
10 Annotations · No Other Contributors

Tell me, O muse, of that ingenious hero who tr … Many cities did he visit, and many were the nations with whose manners and cust … rying to save his own life and bring his men safely home; but do what he might he … folly in eating the cattle of the Sun-god Hyperion; so the god prevented them from … ter of Jove, from whatsoever source you may know them.

So now all who escaped death in battle or by sl … vas longing to return to his wife and country, was detained by the goddess Calypso, … s years went by, there came a time when the gods settled that he should go back to Ithaca; even then, however, when he was among his own people, his troubles were not yet over; nevertheless all the gods had now begun to pity him except Neptune, who still persecuted him without ceasing and would not let him get home.

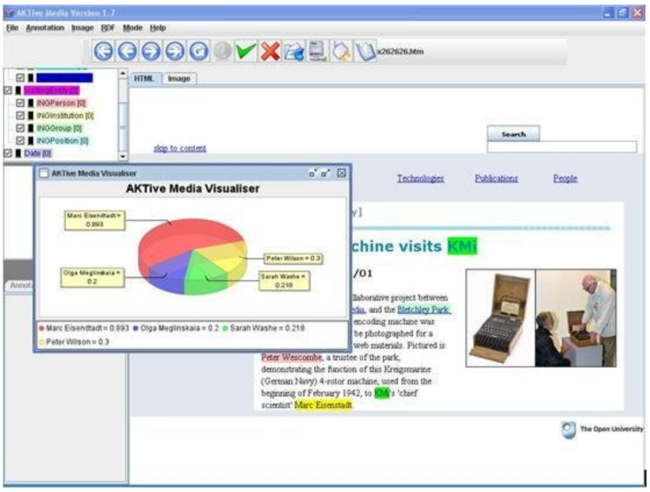
Now Neptune had gone off to the Ethiopians, who are at the world's end, and lie in two halves, the one looking West and the other East. He had gone there to accept a hecatomb of sheep and oxen, and was enjoying himself at his festival; but the other gods met in the house of Olympian Jove, and the sire of gods and men spoke first. At that moment he was thinking of Aegisthus, who had been killed by Agamemnon's son Orestes; so he said to the other gods:

SEMI-AUTOMATIC SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.36. Recogito

supports manual and automatic annotation, where the automatic annotation is proposed with a learning method that explores past annotations made by the user and textual information to make annotation suggestions automatically. GoNTogle provides advanced searching facilities by utilizing a flexible combination of keyword-based and semantic-based searches over the different document formats.

GoNTogle (2010)



chine visits KMi

/01

llaborative project between … dia, and the Bletchley Park … encoding machine was … be photographed for a … web materials. Pictured is Peter Wescombe, a trustee of the park, demonstrating the function of this Kriegsmarine (German Navy) 4-rotor machine, used from the beginning of February 1942, to KMi's 'chief scientist' Marc Eisenstadt.

SEMI-AUTOMATIC SEMANTIC ANNOTATION DIGITAL LIBRARY

Figure 3.37. GoNTogle

3.7 Exploration of a digital library

Many tools face the challenge of exploring the contents of a digital library, but two in particular go in the same direction of this work.

Yewno Discover [7] (fig. 3.38) is an integrated system that offers classification and visual exploration of academic materials to help scholars in their research, but is not adaptable and flexible to different contexts of use, except with ad hoc adjustments. Furthermore, in respect to the proposed system, it makes limited use of the KG structure for exploration, which is at the core of the research questions posed here.

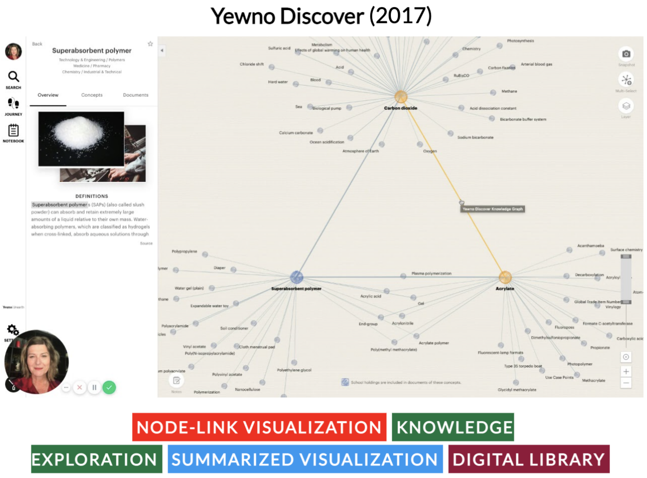


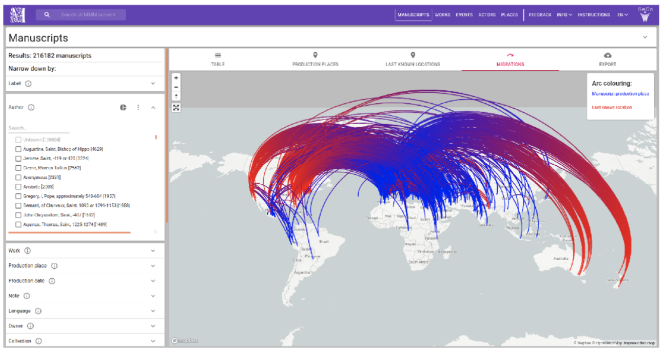
Figure 3.38. Yewno Discover

Sampo-UI [89] (fig. 3.39) is a framework that provides a set of reusable and extensible components, application state management, and a read-only API for SPARQL queries, which can be used to create a user interface for a semantic portal. Sampo uses different search paradigms: free-text search, faceted search, geospatial search, and temporal search. It provides the users with different views of the search results in tables, lists, geospatial, or temporal visualizations. Differently from Sampo UI, the system proposed in this thesis, offer also a knowledge extractor service from unstructured data and a semantic enrichment service.

Another tool used for explore a digital library is Talk to Books. **Talk to Books**[17] is a tool by Google to explore ideas and discover books by getting quotes that respond to user's queries. It aims at helping users to find relevant books that may not be directly identified through keyword search, but does not provide a way for the user to autonomously explore the underlying knowledge base.

[17] https://books.google.com/talktobooks/

Sampo UI (2021)



NAMED ENTITY RECOGNITION TABULAR
VISUALIZATION KNOWLEDGE EXPLORATION MANUAL SEMANTIC
ANNOTATION DIGITAL LIBRARY

Figure 3.39. Sampo-UI

Chapter 4

A visual searching tool for cultural digital artifacts

Contents

4.1 Introduction

In this work, it was analyzed the applicability and usefulness of a corpus search and exploration paradigm based on the transparent use of knowledge graphs. To this purpose, a tool called Arca has been developed. It includes a pipeline for semantic enrichment of textual content and a user interface that enables search and exploration of the corpus through visual navigation of a knowledge graph of topics.

The extracted semantics and user interface is supporting different general search behaviors, which were deemed useful in the specific use case. The main general supported search behaviors are the following:

- find documents relevant to a specific topic;
- expand or specialize searches by moving through related topics;
- have visibility of available related resources, which could potentially be of interest;
- visually organize the resources found by considering their relationships and properties;
- find topics and documents at the crossing of multiple topics, possibly of different kinds (places, people, time periods, etc.).

4.1.1 Comparison with the SOTA

The analysis of tools (chapter 3) for the management and visualization of semantic data, tagged by characteristics, functionality and paradigms of visualization and interaction (fig. 3.2, 3.3), leads to the presentation of the tool proposed in this thesis: Arca, the acronym of Academic Research Creativity Archives (fig 4.1).

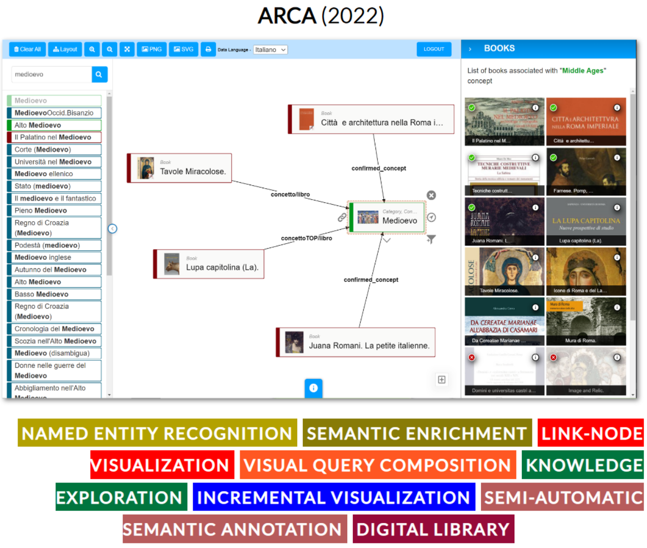


Figure 4.1. Arca tool

Like the L'Erma and Torrossa tools, Arca was created to search and view a digital library's contents. Unlike the latter, Arca is a modular tool (like Apache Stanbol) that includes a knowledge extraction and semantic enrichment engine and an interface for searching and exploring data. The graph-based visualization has been selected as a paradigm with a multi-level representation of information. The books and their associated contents are highlighted at the centre of everything. The mode of interaction is incremental concerning the information sought: we pass from the particular to arrive at the general. Interaction supports serendipity, a concept according to which unexpected information can be reached by exploring and navigating the graph. Arca allows the creation of visual queries through the component "trace path" integrated. Trace path allows the achievement of information common to two selected concepts, such as books containing two concepts or concepts common to two books. Furthermore, Arca integrates an association validation system to support a collaborative improvement of data quality.

An in-depth analysis of the tool proposed in this thesis is presented in the next chapter 4.

In conclusion, we can summarize the research contribution that the development of the Arca tool (protagonist of this thesis) has brought concerning the state of art.

Research contribution

Compared to the tools analyzed so far, Arca allows the incremental exploration of knowledge graph information (supporting the principle of serendipity) with the novelty of application to the exploration context of a digital library. Arca aims to facilitate the search, viewing and exploration of books or documents that deal with the information sought. Among the tools analyzed, no system focuses on exploring a catalogue of books that exploits the interaction paradigm adopted by Arca.

The remaining sections are organized as it follows. Section 4.2 reports the design process starting from identifying user requirements to the final interface's development and implementation. Section 4.3 introduces some technical background needed for appreciating the proposed system, which is then detailed in Section 4.4. Section 4.5 reports the evaluation process and analyzes the findings.

4.2 System requirements

The publishing house's specific use case offered the opportunity to adopt a user-centered design approach to identify and refine the system requirements. From informal interviews with publishing house representatives and a team of researchers in the same domain an initial set of requirements has been identified:

- the user should be able to search entities textually;
- for an entity, the user should be able to see the relevant books;
- the user should be able to navigate among entities, following semantic relationships between them;
- for a document, the user should be able to access the basic information and be informed on how to obtain it (buy it from a bookstore, borrow it from a library, etc.);
- any user should be able to perform operations without being taught how to, by following established interaction patterns and metaphors.

A series of intermediate evaluations were carried on, including the following methods:

- evaluation of extracted data quality by expert analysis;
- tests and discussion with low fidelity prototypes (as an example, the reader can see the mockup in Figure 4.2, which was one of the initial proposal, and can compare with the interface shown in Figure 4.7);
- tests and discussion with high fidelity prototypes (progressively closer to the final system).

During these iterations, the following additional requirements were identified:

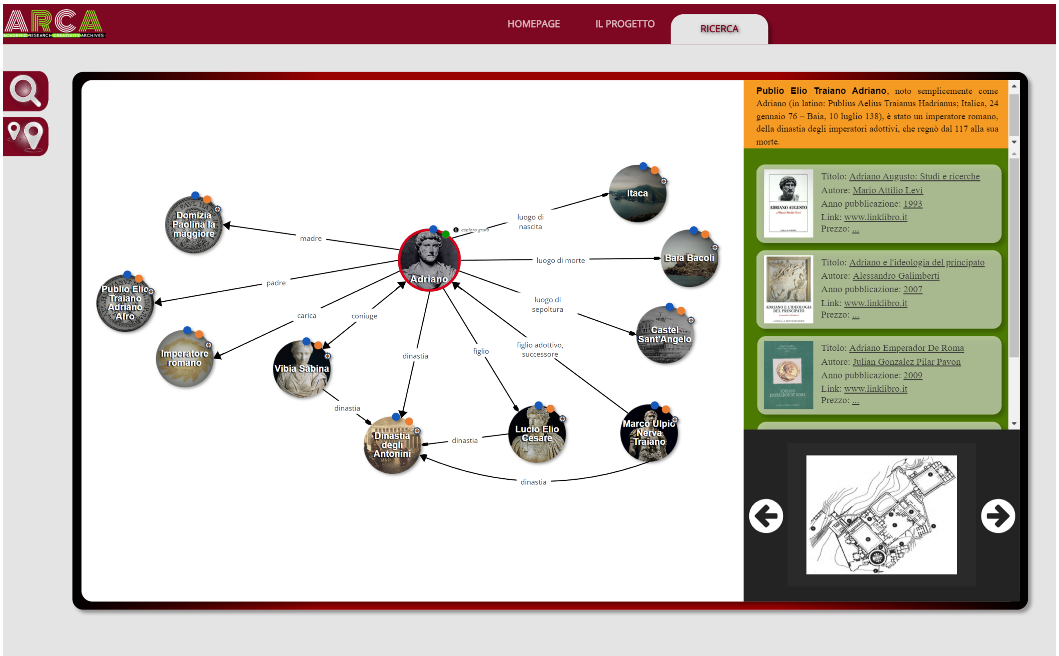
- entities which appear more frequently in a document (main topics) should be distinguished from less relevant entities;
- users need to check the textual context in which an entity was found in a document.

4.2.1 Approach

Arca is the software system designed to enable the KG-based exploration of a given text corpus and test the listed hypotheses.

The system is organized according to the following main functions:

- extraction of entities from a given text corpus;
- integration between available metadata, extracted entities present in the text, and data from external knowledge bases;



Publio Elio Traiano Adriano, noto semplicemente come Adriano (in latino: Publius Aelius Traianus Hadrianus; Italica, 24 gennaio 76 – Baia, 10 luglio 138), è stato un imperatore romano, della dinastia degli imperatori adottivi, che regnò dal 117 alla sua morte.

Titolo: Adriano Augusto: Studi e ricerche
Autore: Mario Attilio Levi
Anno pubblicazione: 1993
Link: www.linklibro.it
Prezzo: ...

Titolo: Adriano e l'ideologia del principato
Autore: Alessandro Galimberti
Anno pubblicazione: 2007
Link: www.linklibro.it
Prezzo: ...

Titolo: Adriano Emperador De Roma
Autore: Julian Gonzalez Pilar Pavon
Anno pubblicazione: 2009
Link: www.linklibro.it
Prezzo: ...

Figure 4.2. The mockup of the user interface.

- consolidation of the local data in a KG stored in a triple store;
- search and exploration of the corpus through the navigation of the KG in a composite user interface.

In order to ensure the whole solution is useful for potential users, it has been implemented and evaluated within a specific case study: exploring the book catalog of a medium-sized publishing house specializing in ancient history. The concrete case study offered the context for fruitful exchange among the stakeholders that are often involved in scenarios of information retrieval and library search:

- who maintain the corpus (the publisher);
- who need to search the corpus (researchers of the field and interested individuals);
- who develop the software solution (in this case the authors of the present study).

4.2.2 Research questions

For the sake of the analytic approach, it was framed the experimentation effort through a set of research questions. The questions elicited below are relevant to the application of KG-based approaches for the exploration of text corpora.

Q1 Would users exploring a corpus of text profit from the semantic navigation of the associated KG of topics?

Q2 What kind of user interface would effectively support such a navigation?

Q3 What kind of users, scenarios, and tasks would benefit from this interaction paradigm?

Q4 Does building and maintaining a semantic enrichment and KG creation pipeline necessarily involve high upfront costs and highly skilled developers?

4.2.3 Hypotheses

To reply to the questions above, the presented study it was designed to test the following main hypotheses.

H1 (relevant to Q1 and Q2). Users will be able to effectively explore a text corpus through a KG-based user interface, which offers the following main functions: *a.* finding concepts through text search (among the ones pertinent to the specific domain), *b.* visually navigating the concepts and their relationships, and *c.* showing documents relevant to the selected concept.

H2 (relevant to Q3). The method, given a corpus of texts in a specific domain, will benefit both users with little knowledge of the domain (by supporting semantically-relevant discovery) and domain experts (by enabling a topic-oriented visual organization of the documents).

H3 (relevant to Q4). It is feasible to build a ready-to-use complete system, including both semantic enrichment pipeline and web-based front end, which is able, with only some configuration, to be applied to any specific corpus to enable the KG-based exploration.

While the first two research questions and related hypotheses are relevant for investigating the benefits of the proposed approach for the end users, the last research question and hypothesis investigate the usefulness and portability of such a system to different contexts of use.

4.3 Technical background

This section describes the technologies underlying the proposed system briefly. Semantic technologies enable the transformation of unstructured information, like those present in the textual PDF documents, to structured data.

The semantic web [31, 32], according to Berners-Lee, is “a web of things of the world, described by the data on the web” [33]. The concept is generic, but contains some crucial references:

- the network (graph);
- things (objects related in a meaningful way);
- data (no longer records, but connections among nodes of a network).

The concept of the semantic web is closely related to the concept of linked data as an effective method and technique for simplifying and homogenizing solutions to identity and interoperability problems, promoting the univocal identification of data in the dialogue between heterogeneous systems.

Linked data [146] is based on a set of techniques that, through shared vocabularies, allows non-human agents to understand content published on the web. The linked data initiative includes both technology and a set of best practices for publishing data on the web in a readable, interpretable, and usable way by a machine.

A knowledge graph [147] (section 2.3) is a set of graph-structured data where nodes represent entities of interest and edges represent relationships between them. A part from the data model, what distinguishes a knowledge graph from a typical database is that it is not tied to a specific application. A knowledge graph is meant to hold information that is of interest for a company, a community, a domain of knowledge, or even include data from multiple domains. This graph is often enriched with various forms of schemata, rules, ontologies, and more, to help validate, structure, and define the semantics of the underlying graph. When considered at web scale, the idea of knowledge graphs overlaps with the concept of linked data.

Making data understandable to machines implies the sharing of a typical data structure. The RDF (Resource Description Framework) [34] is the language proposed by the W3C for achieving a standard data structure as a graph, in the context of linked data. In RDF data are organized around *resources*. Relations between resources are represented through *triples*, i.e. subject-predicate-object associations. Subject and object are a pair of related resources. The predicate is another resource which specify the meaning of the relation. Resources used in the predicate role are called *properties*. Furthermore, the object of a triple can also be a *literal*, i.e. a simple value conforming to some basic datatype as string, integer, date, etc. Resources have types too, which are specific resources called *classes*. A resource may have multiple types. An *RDF graph* is defined as a set of triples.

SPARQL [151] is one of the key technologies of the semantic web, and it is used to retrieve and manipulate RDF data from the knowledge graphs available on the web. The SPARQL endpoints allow clients to issue SPARQL queries over a dataset, receiving direct results from the server.

4.4 The system

The software system has been implemented and tested in the specific use case, but it is designed for general use. The aim is to offer a ready-to-use package to explore any corpus of texts through a specialized KG visually. In this section, the software is described starting from its modular organization. Then an overview of the data model and structure of the integrated KG is presented. Later the user interface is described. Finally, the implementation details are given.

4.4.1 Software modules

Figure 4.3 shows informally the system's main software modules, the different user categories, and the data flow among them. For clarity, modules and flows are organized in the main functional areas (numbered from 1 to 4). The system is composed of a pipeline to build the KG and a web-based front end to search the corpus using the KG.

The pipeline can be seen as composed of three steps, roughly corresponding to the functional areas 1, 2, and 3 of the diagram in Figure 4.3: in the first step, newly added documents of the corpus enter the pipeline; in the second step, semantic enrichment services extract information from the documents; in the third step, the generated data is consolidated locally to be also integrated with additional data provided by external services.

RDF is used to represent all the data items in the pipeline, employing existing vocabularies and ontologies whenever possible and creating new terms if needed.

In the first step, the documents' content (e.g., PDFs) is stored in the system, along with the relevant metadata. In the current version, the documents are loaded by copying them in a directory, but it will be generalized with a repository that supports the linked data container API. The repository will then be maintained by the catalog maintainers (e.g., editors or librarians) through a dedicated front-end application. It will also be possible to connect it directly with existing databases or systems for automatic content insertion or update.

In the second step, the documents analyzed by a set of semantic enrichment services give as output the knowledge extracted from the content expressed using existing models and KGs. Currently, a service providing entity extraction (Dandelion Api[1]) is called. The result is a set of the recognized entities (identified as DBpedia[2] resources) alongside the document's point in which the entity was found. An adapter converts this information to RDF to be later integrated with existing metadata and the DBpedia KG.

[1] Dandelion API - `https://dandelion.eu/`

[2] `https://wiki.dbpedia.org/`

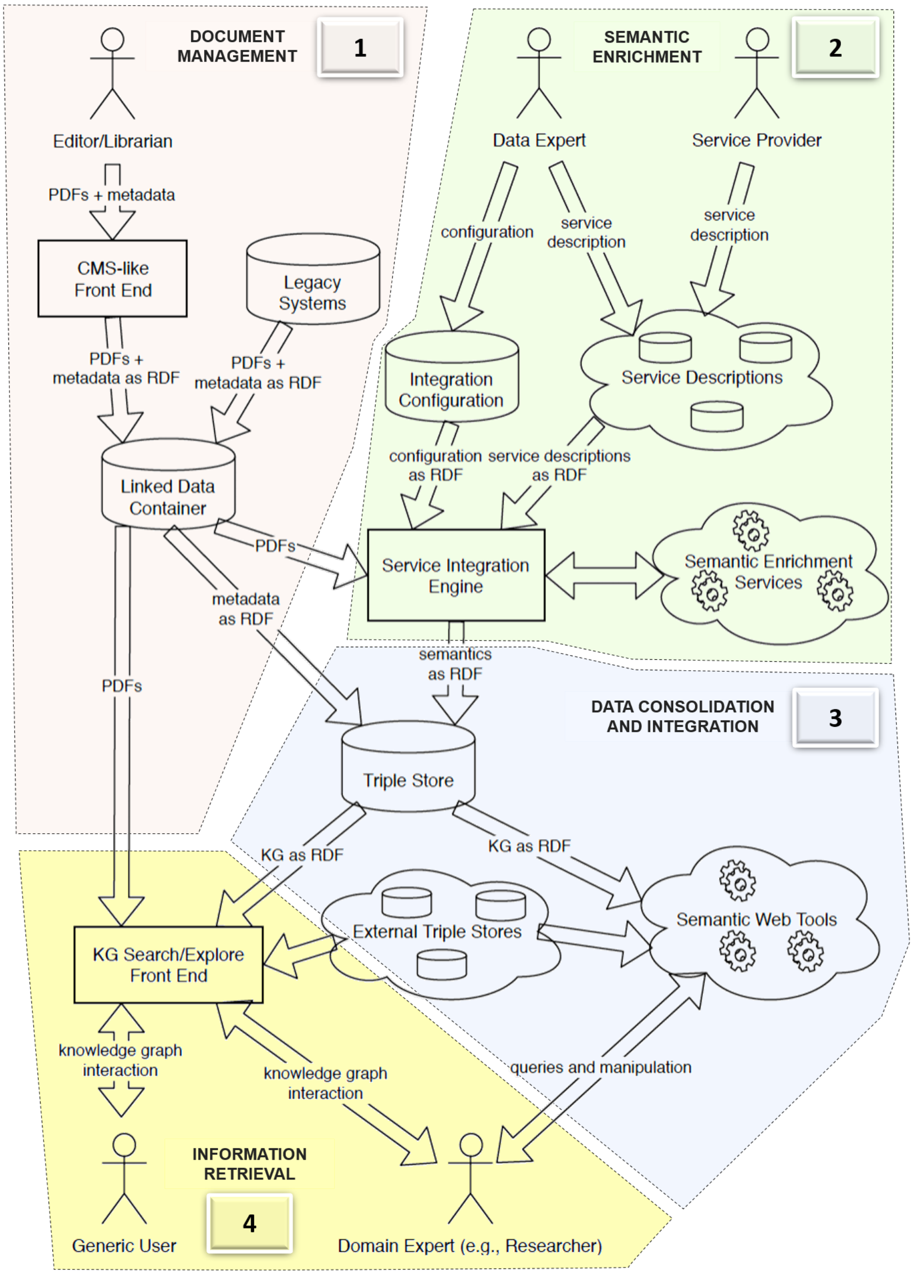


Figure 4.3. Diagram describing the flow of data in the system.

In the third step, both the metadata coming from the linked data container and the knowledge extracted in the previous step are stored in a triple store as an integrated KG. Due to the distributed nature of linked data, relevant additional external data may be either added to the triple store in this step or kept separated and accessed on demand when needed.

Finally the “Information Retrieval” functional area (number 4 in Figure 4.3) refers to the actual usage of the KG to search and explore the corpus by generic users as well as domain experts. Users can use a web-based front end offering the visual user interface described in Section 4.4.3. The front end is able to integrate on the fly data from the local triple store and other linked data sources. In the specific use case, the data from DBpedia is integrated on demand, so that the explored KG is a virtual graph obtained by merging the local KG with the DBpedia KG. Furthermore, the local data being in a triple store, direct access through a SPARQL endpoint can be enabled, thus providing expert users with a means to perform advanced queries and further integration.

4.4.2 The data model

The data gathered in the process described in Section 4.4.1 is stored as a knowledge graph, which will be referred to as Arca Knowledge Graph.

The Arca KG is an RDF graph that describes:

- information extracted automatically during the knowledge extraction process from books;
- existing metadata associated with books.

The Arca KG makes use of multiple vocabularies, providing a set of classes and properties to describe the given domain.

Tables 5.1 and 5.2 list respectively classes and properties employed to define Arca KG. Listing 4.4 shows a fragment of RDF describing some of the information associated with a book.

The data model incorporates a new vocabulary, described below, as well as the following existing vocabularies.

SKOS [3] is a common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems, and taxonomies. The property `skos:broader` is adopted to define hierarchical relations among concepts. See lines 54 and 58–61 of Listing 4.4.

FOAF[4] provides terms for describing people and organizations, documents associated with them, and social connections between people. The property `foaf:depiction` is used to associate the books with their cover image. See line 23 of Listing 4.4.

SCHEMA[5] provides to mark up website content with metadata about itself. Properties from SCHEMA are used to connect the authors and the ISBN codes to the books. See lines 7 and 9 of Listing 4.4.

[3] see `https://www.w3.org/TR/swbp-skos-core-spec/`

[4] Friend Of A Friend - see `http://xmlns.com/foaf/0.1/`

[5] Schema.org - see `https://schema.org/`

Table 4.1. Arca KG Classes

name	**description**
FROM ARCA NAMESPACE	
arca:Book	A "Book" is a PDF document analyzed and inserted into the Arca system. The use case included books on ancient Roman history, but the ontological skeleton described here is applicable to any subject area.
arca:Snippet	A "Snippet" is a text fragment which represents the textual contexts of the document in which a concept has been found
FROM OTHER NAMESPACES	
dbo:Person	A person (alive, dead, undead, or fictional).
madsrdf:Temporal	Describes a resource whose label represents a time-based notion.
madsrdf:Topic	Describes a resource whose label represents a topic.
madsrdf:GenreForm	Describes a resource whose label is a genre or form term. For example, biographies, catechisms, essays, hymns, or reviews, daybooks, diaries, directories, journals, memoranda, questionnaires.

MADS/RDF (Metadata Authority Description Schema in RDF)[6] is a data model for authority and vocabulary data used within the library and information science (LIS) community, which is inclusive of museums, archives, and other cultural institutions. Here classes from MADS/RDF have been adopted to identify different classifications of books present in meta-data. See lines 44, 47, and 50 of Listing 4.4.
DC (Dublin Core)[7] is a metadata vocabulary used by many libraries. Properties from DC are used to connect the books' title, date of publication, language and abstract to the books. See lines 3–6 of Listing 4.4.
DBO (DBpedia Ontology)[8] is the ontology used within DBpedia. Here Person class from DBO has been adopted to define books' authors. See lines 35, 38, and 41 of Listing 4.4.

Documents are defined by the Arca class `arca:Book` (cfr. Table 5.1; line 1 of Listing 4.4). The metadata concern:

- the title, language, publication date, and abstract of each book (lines 2–6);
- the authors (see `schema:author` lines 10–12, 35, 38, and 41);
- the topic of the book (see `dc:subject` lines 15 and 44–45);
- the type of book (see `dcterms:type` lines 18 and 47–48);

[6] MADS/RDF - `https://id.loc.gov/ontologies/madsrdf/v1.html`

[7] Dublin Core - see `https://dublincore.org/`

[8] DBpedia Ontology - see `https://www.dbpedia.org/resources/ontology/`

Table 4.2. Arca KG Classes Properties

name	description	domain	range	cardinality
FROM ARCA NAMESPACES				
arca:concept	Represents the concept of a document	arca:Book	owl:Thing	multiple
arca:top_concept	Represents the top concept of a document	arca:Book	owl:Thing	max(10)
arca:containEntity	Refers to the snippet for the extracted concept	arca:Snippet	owl:Thing	funcitonal
arca:intoBook	Refers to the document in which the snippet appears	arca:Snippet	arca:Book	functional
FROM OTHER NAMESPACES				
rdfs:label	Represents the human-readable name of a given resource	rdfs:Resource	rdfs:Literal	functional
dc:date	A point or period of time associated with the publication date of the document	rdfs:Resource	rdfs:Literal	functional
dc:language	A language of the document	rdfs:Resource	dbo:Language	functional
dc:title	The title of a document	rdfs:Resource	rdfs:Literal	functional
dc:about	further specifications on the document	rdfs:Resource	rdfs:Literal	functional
dc:subject	A topic of the document	arca:Book	madsrdf:Topic	functional
dc:description	The description used to specify the content of a Snippet	arca:Snippet	xsd:string	functional
dcterms:type	The genre of the document	arca:Book	madsrdf:GenreForm	functional
schema:ISBN	The ISBN of the book	rdfs:Resource	xsd:string	functional
schema:author	The author of the document	arca:Book	dbo:Person	multiple
dcterms:temporal	Temporal characteristics of the document	arca:Book	madsrdf:Temporal	functional
foaf:depiction	Represents the cover image of a book	arca:Book	foaf:Image	functional
skos:broader	Explicit expression of the relationships between and among concepts	arca:Book	owl:Thing	multiple

- the era of which the book narrates (see `dcterms:temporal` lines 21 and 50);
- the cover of the book (see `foaf:depiction` line 23);

```
 1 lermabook:DE000059 a arca:Book ;
 2             rdfs:label "Ara Pacis Augustae." ;
 3             dc:title "Ara Pacis Augustae." ;
 4             dc:language "it" ;
 5             dc:date 1986;
 6             dc:about "In occasione del restauro della fronte orientale." ;
 7             schema:ISBN "9.78888E+12" ;
 8
 9             schema:author
10                 author:la_rocca_eugenio,
11                 author:zanardi_bruno,
12                 author:ruesch_v;
13
14             dc:subject
15                 metadata_subject:restauro_conservazione ;
16
17             dcterms:type
18                 metadata_type:mostre_cataloghi ;
19
20             dcterms:temporal
21                 metadata_age:eta_classica ;
22
23             foaf:depiction lermabookimg:DE000059.jpg ;
24
25             arca:concept
26                 <http://dbpedia.org/resource/Aeneas> ,
27                 ...
28                 <http://it.dbpedia.org/resource/Roma_(citt%C3%A0_antica)> ;
29
30             arca:top_concept
31                 <http://dbpedia.org/resource/Augustus> ,
32                 ...
33                 <http://dbpedia.org/resource/Altar> .
34
35         author:la_rocca_eugenio a dbo:Person ;
36             rdfs:label "La Rocca Eugenio" .
37
38         author:zanardi_bruno a dbo:Person ;
39             rdfs:label "Zanardi Bruno" .
40
41         author:ruesch_v a dbo:Person ;
42             rdfs:label "Ruesch V." .
43
44         metadata_subject:restauro_conservazione a madsrdf:Topic ;
45             rdfs:label "restauro conservazione" .
46
47         metadata_type:mostre_cataloghi a madsrdf:GenreForm ;
48             rdfs:label "mostre cataloghi" .
49
50         metadata_age:eta_classica a madsrdf:Temporal ;
51             rdfs:label "eta classica" .
52
53         <http://dbpedia.org/resource/Aeneas> a arca:Concept ;
54             skos:broader <http://dbpedia.org/resource/Kings_of_Alba_Longa> ;
55             rdfs:label "Enea" .
56
57         <http://dbpedia.org/resource/Symbol> a arca:Concept ;
58             skos:broader <http://dbpedia.org/resource/Semiotics> ;
59             skos:broader <http://dbpedia.org/resource/Semiotica> ;
60             skos:broader <http://dbpedia.org/resource/Communication_design> ;
61             skos:broader <http://it.dbpedia.org/resource/Sociologia_della_cultura> ;
62             rdfs:label "Simbolo" .
63
64         snippet:DE000059_SNI1 a arca:Snippet ;
65             rdfs:label "Augusto";
66             dc:description "D'altronde lo stesso Augusto era conscio della sua posizione";
67             arca:containEntity
68                 <http://dbpedia.org/resource/Augustus>;
69             arca:intoBook
70                 <http://www.lerma.it/index.php?pg=SchedaTitolo&key=DE000059>.
```

Figure 4.4. RDF Fragment.

The information extracted automatically concerns all the concepts contained in each book (see `arca:concept` line 25) and the main ten concepts that describe a book (see `arca:top_concept` line 30) together with the position of the character of beginning and end in the text where the concepts were extracted. This last information is used to generate *snippet* resources (lines 64–70), having type `arca:Snippet` and associated to a text context (`dc:description` line 66) of the extracted concept (`arca:containEntity` line 67) from a specific book (`arca:intoBook` line 69).

Four dedicated namespaces have been defined to build unique IRIs from categories in the existing metadata. In the Listing 4.4 they are associated with the following prefixes:

- `author:` for authors;
- `metadata_subject:` for topic;
- `metadata_type:` for book genres;
- `metadata_age:` for historical time periods.

Figure 4.5 shows some of the properties described in the RDF fragment, as part of the user interface described later. Figure 4.6 is another view of the user interface, in which part of the RDF graph described in the fragment is represented visually. The `arca:Snippet` class, described in Table 5.1, is illustrated in Figure 4.8.

4.4.3 The user interface

The visual user interface[9] is composed of two main components (see Figure 4.7). The first component contains the visualization and search of the entities (part 1 fig. 4.7) contained in the KG. It is a customized version of the Ontodia workspace (briefly described in Section 3.5). The second component shows the list of documents associated with the selected entity (part 2 fig. 4.7), offering further interaction.

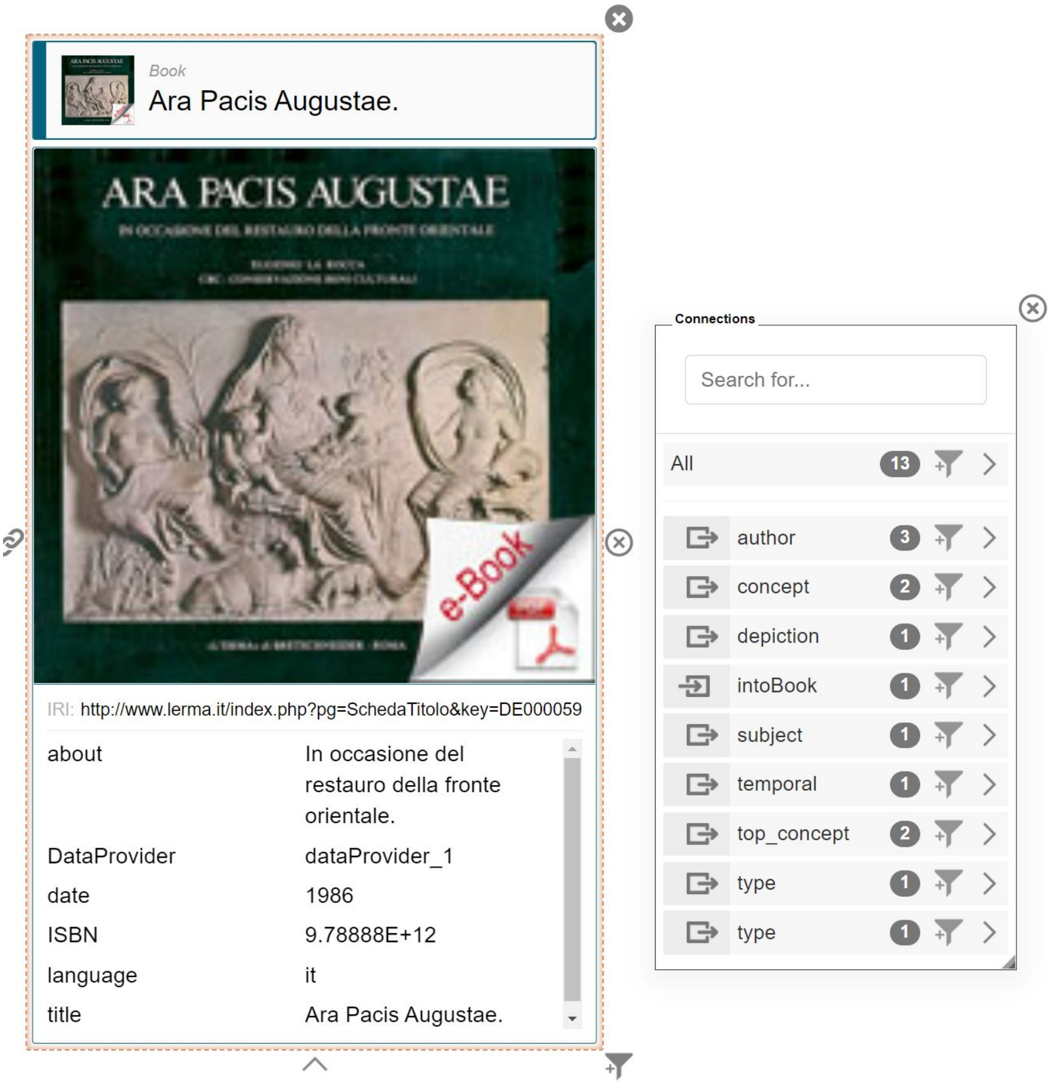
Exploration of the knowledge graph

The knowledge exploration component (see part 1 of Figure 4.7) has the following features.

Searching graph entities. The left panel enables search for entities in the knowledge graph, corresponding in the use case mainly with entities from DBpedia. For example, typing "Rome" the user gets all the entities containing that string in their label. One or more of the returned entities (e.g., the one corresponding to Rome's city) may be loaded to the graph navigation panel through drag-and-drop.

Knowledge graph navigation. The central panel allows the user to navigate the KG. Starting from any shown entity, its connections, i.e. RDF triples in which the given entity is subject or object, can be expanded (hence adding the connected

[9] `http://arca.diag.uniroma1.it:5000`



Book

Ara Pacis Augustae.

IRI: http://www.lerma.it/index.php?pg=SchedaTitolo&key=DE000059

about	In occasione del restauro della fronte orientale.
DataProvider	dataProvider_1
date	1986
ISBN	9.78888E+12
language	it
title	Ara Pacis Augustae.

Connections

Search for...

All	13
author	3
concept	2
depiction	1
intoBook	1
subject	1
temporal	1
top_concept	2
type	1
type	1

Figure 4.5. Visual exploration of a resource properties 4.4

entities to the graph). Rather then expanding all the connections the user may select specific RDF properties (e.g., *birthplace*). Figure 4.5 shows the box with expanded information about a book along with the box to chose connections by RDF property. Furthermore, the connections among shown entities are shown by default, as they may be of interest. The navigation panel is coordinated with the document list panel (described below and shown in part 2 of Figure 4.7) so that the latter shows the list of documents which include as topic the entity currently selected in the former.

Documents as entities. Apart from being shown in the document list panel, documents can also be explored as entities themselves in the KG exploration.

They are linked to their topics by two types of semantic connections: *concept* for

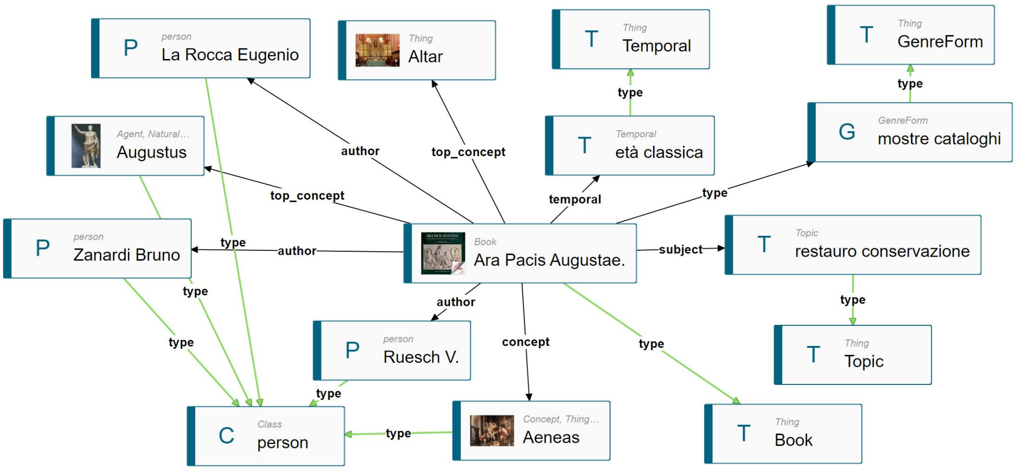


Figure 4.6. Visual exploration of the RDF fragment in Listing 4.4

any entity found in the text, *top concept* for the ones recognized as main topics for that text. This choice enables further ways to interact with the system:

- starting from a document, to explore its topics and then possibly other documents from them (e.g., in Figure 4.7, from the book *The Tale of Cupid and Psyche* to the topic *Rome* and then to the book *Scutulata Pavimenta*);
- from shown entities, to visualize which documents are about two or more of them (e.g., in Figure 4.7, the book *The Tale of Cupid and Psyche* is both about *Rome* and, specifically, about *Castel Sant'Angelo*).

Kinds of entities. Different colors are used as an aid to distinguish three broad sets of entities:

- DBpedia entities not found in the corpus are in blue;
- DBpedia entities found at least once in the corpus are in green;
- documents are in red.

Document list

The document list panel (part 2 of Figure 4.7), which can be shown or hidden as needed, shows the list of documents associated with the entity currently selected in graph exploration panel, i.e. the documents whose extracted entities include that one. The documents may be shown ordered by year of publication or by relevance (if for that document it is a *main topic* or just a *topic*). By clicking on the *info* button of a book, a modal window with further information on the document is

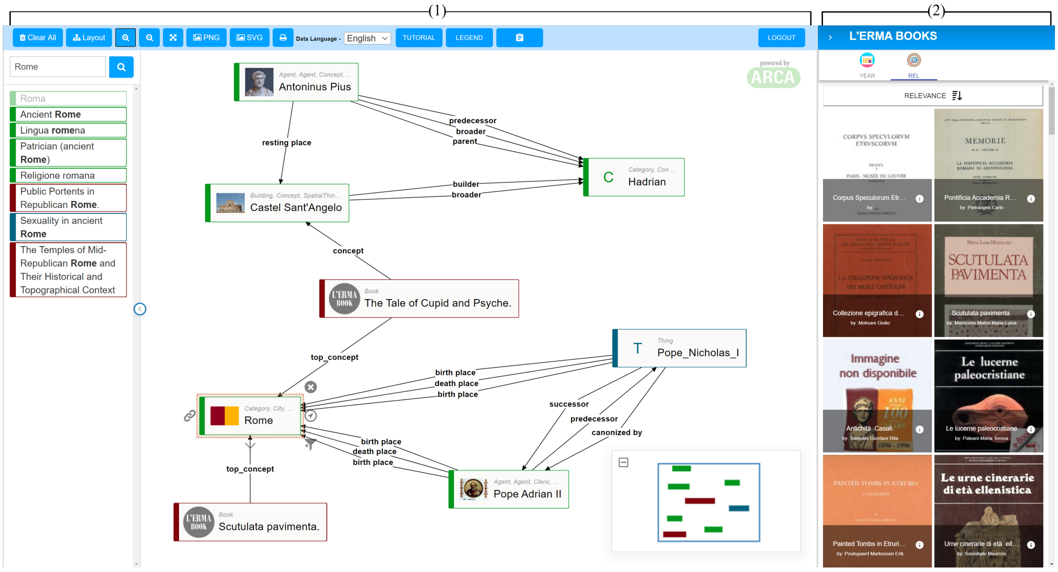


Figure 4.7. The user interface.

opened (see Figure 4.8). The information includes the list of *snippets*, i.e., all the textual contexts of the document in which the selected concept has been found.

Trace path

The tool shows the connections between two selected entities. Thanks to the complex queries that can be processed on the knowledge graph, this tool identifies all the books connected to the first selected entity (in the case of Figure 4.9 "Ancient Rome") and all the books connected to the second selected entity ("Monument") and makes an intersection on the two sets, showing only the books and links in common.

4.4.4 Implementation

For the triple store at the core of the system, it is Blazegraph[10], while the pipeline which builds the KG is developed in Python.

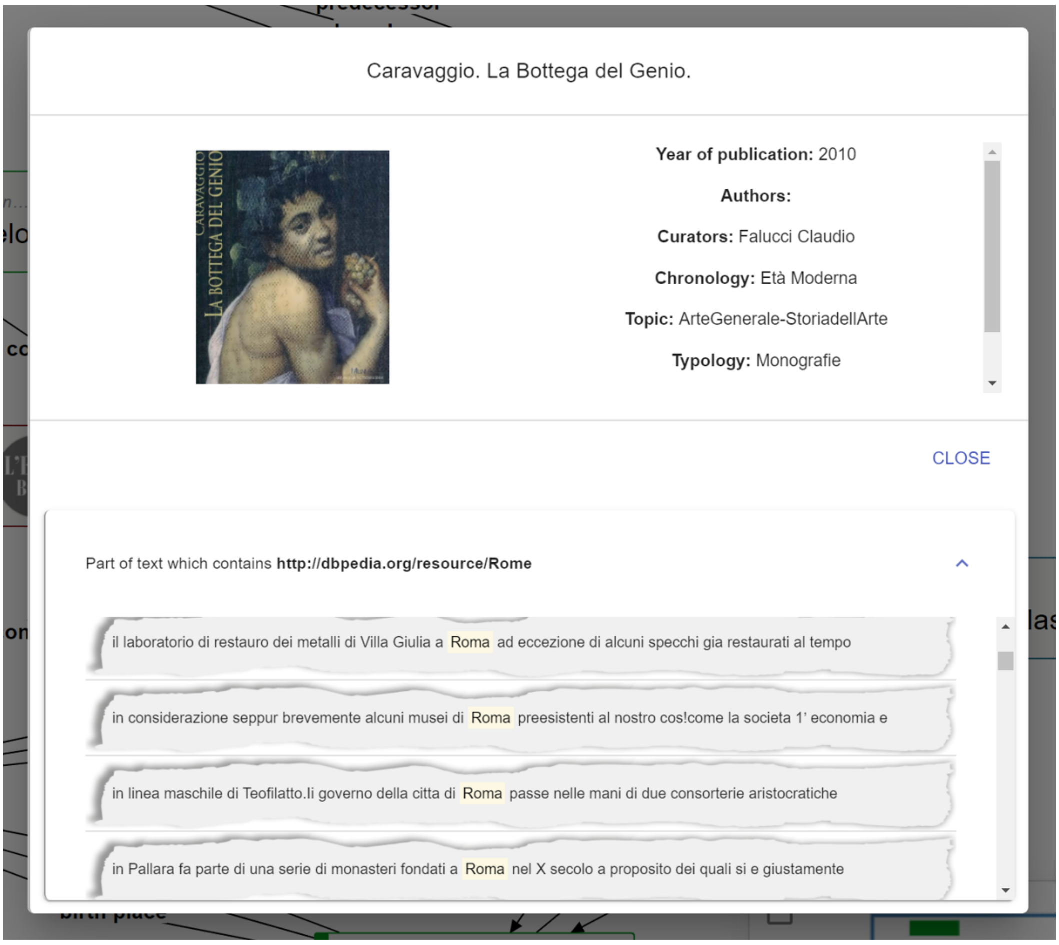
The web front end is developed using the React framework[11] for modularity. It includes customized components from Ontodia as well as components built from scratch. The code is maintained in a public repository on GitHub[12].

In the publishing house use case, some books do not exist natively in electronic format, for which the scanned pages go through the OCR of the software ABBYY

[10] https://www.blazegraph.com/

[11] https://reactjs.org/

[12] https://github.com/EleonoraAI/ARCA-frontend



Caravaggio. La Bottega del Genio.

Year of publication: 2010

Authors:

Curators: Falucci Claudio

Chronology: Età Moderna

Topic: ArteGenerale-StoriadellArte

Typology: Monografie

CLOSE

Part of text which contains **http://dbpedia.org/resource/Rome**

il laboratorio di restauro dei metalli di Villa Giulia a Roma ad eccezione di alcuni specchi gia restaurati al tempo

in considerazione seppur brevemente alcuni musei di Roma preesistenti al nostro cos!come la societa 1' economia e

in linea maschile di Teofilatto.li governo della citta di Roma passe nelle mani di due consorterie aristocratiche

in Pallara fa parte di una serie di monasteri fondati a Roma nel X secolo a proposito dei quali si e giustamente

Figure 4.8. Sentences.

FineReader Pro 15[13]. For semantic enrichment, it was used the external *entity extraction* (NERL) web service offered by the Dandelion API[14]which offers several text analysis services for many languages: entity extraction, text similarity, text classification, language detection and sentiment analysis. Dandelion relates segments of the input text to resources in Wikipedia, along with a confidence value. Nevertheless, given the flexibility of the Arca service integration mechanism, the system is not tied to this specific service.

4.4.5 Choices and motivations

In Section 4.4, the sub-elements of the system have been described in detail. In this final subsection, were deepen the purposes that led to the design and development of each system element.

[13] https://www.abbyy.com/en-eu/finereader/

[14] https://dandelion.eu/

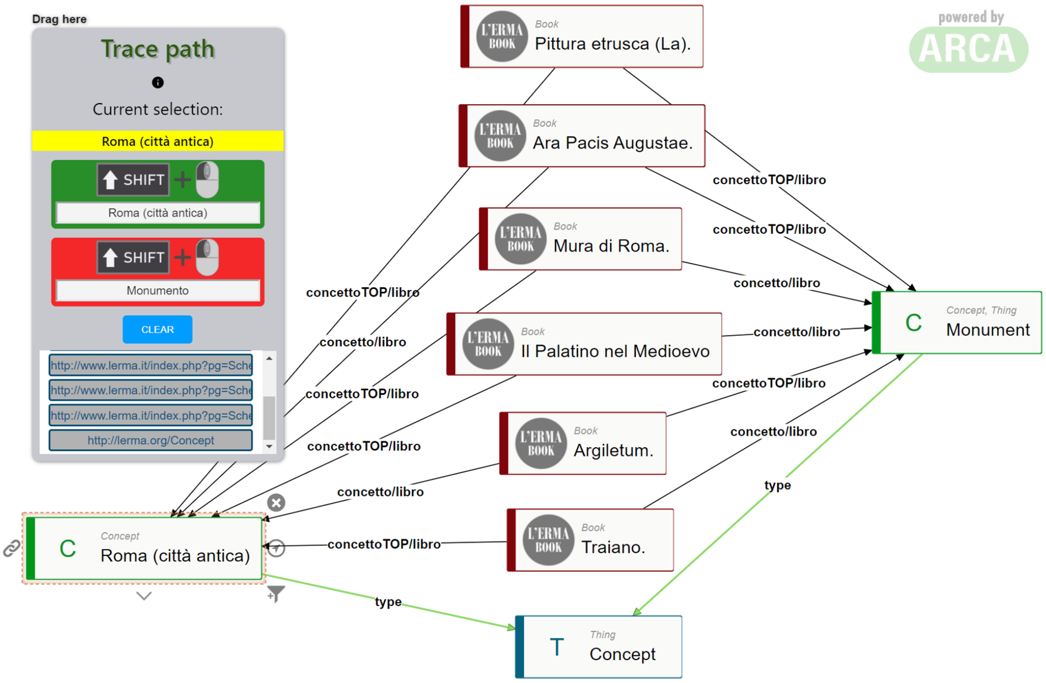


Figure 4.9. Trace path.

Following, are listed the choices adopted in the development of the system and its sub-elements that support the motivations and the requirements described in Section 4.2.

- the system's modular design (cf. Section 4.4.1 *Software Modules*), which allows easy management and maintenance of a ready-to-use system. Thanks to the modularity, each module independently manages the different phases of knowledge extraction, semantic enrichment, entity linking, connection to other knowledge bases and complex queries called up through intuitive and usable interface components.
- the search bar (cf. Section 4.4.3 *Searching graph entities*), which allows to search for a topic and to:
    - find concepts consistent with what is being researched and extrapolated directly from the documents of the corpus of texts inserted in Arca;
    - find documents whose title is semantically consistent with what is sought (cf. Section 4.4.3 *Documents as entities*);
    - find other semantically coherent resources deriving from the knowledge graphs integrated into the search system (cf. Section 4.4.3 *Kinds of entities*);

- the graph navigation mode (cf. Section 4.4.3 *Knowledge graph navigation*), which supports the user in accessing connected resources and discovering new information following the philosophy of serendipity.
- direct access to the list of documents (belonging to the digital library included in Arca) in which the topic of interest is discussed.
- the possibility of finding topics in common with different resources (cf. Section 4.4.3 *Trace path*).

4.5 Evaluation

The system has been tested in the context of a specific use case: exploration of the book catalog of medium-size publishing house, specialized in classical antiquity. The anticipated final users of the tool can be roughly classified in two categories:

- domain experts who may adopt a new approach to search and discover resources in the context of their research;
- curious people who want to explore new topics.

Arca's evaluation process lasted two years and was characterised by three phases:

- an evaluation of the extracted data, from the point of view of quality and usefulness, with the help of domain experts;
- a small-scale qualitative user-based evaluation of the tool with a some researchers of the field;
- a larger and richer user-based evaluation of the tool, both on its own and in comparison with other existing solution, which involved both students and researchers of the field.

In the first two evaluation phases, discussed in previous work [82], the focus was on analyzing the limits and margins for improvement of Arca with the involvement of researchers experts in the relevant domain, who had also participated in the design process. It was possible to identify problems in the extraction of topics from the books and get feedback to improve the whole system.

The third evaluation phase involved thirty users and included both a comparative evaluation with other three tools offering similar functionality (Lerma, Torrossa, and Yewno) and a specific evaluation of Arca on its own. Both parts of the evaluation are task-based and contain questions designed to evaluate multiple factors of the user experience and elicit perceived strengths and limitations of the tool. The following subsections describe in detail this experiment: the setup, the obtained results, and their discussion. The raw data gathered from the test are also publicly available online on Zenodo [152].

4.5.1 Setup

As anticipated, the third phase of evaluation has been a user test involving thirty users, who were students and researchers in the field of classical antiquity. Participants had different levels of academic education: nine secondary school qualification, five bachelor's degree, seven master's degree, seven PhD[15].

Choosing students and researchers of the specific field considered in the use case was crucial to the experiment: it allowed to assume at least some level of interest for the considered topics and provided a way to partially predict the level of relevant background knowledge based on the reached academic qualification [16]. Albeit all the involved users study the field at an academic level, their diverse level of academic education partially covers the requirement of H2 to test the tool with users of varying levels of knowledge of the domain.

In order to carry on a comparative evaluation, a task-oriented setup planned. This task enabled the users to access four different tools on equal grounds, in random order, and remaining unaware of the fact that one of the tools (Arca) is developed by us. The comparative evaluation included two tools providing simple text-based search (Lerma and Torrossa) and two tools providing search enhanced by semantics (Yewno and Arca).

The task-oriented comparative evaluation was compounded by a part of the evaluation focusing on specific aspects of Arca. This part was scheduled in the end, to preserve the fairness of the comparative evaluation.

Just when was planed the last phase of evaluation, the COVID-19 pandemic broke out and it was not possible to carry out the third phase of the evaluation in presence. For this reason, the process was re-designed to make users autonomous and able to perform the required activities and answer questions while using the system from home.

Given the richness and complexity of the evaluation, the goal of comparing multiple tools, and the necessity for the users to be able to perform the whole process autonomously, a lot of effort was put in a carefully designed user interface that

[15] Two replied *other*.

[16] We are aware the relationship between education and level of knowledge of a field is far from straightforward, but other measures have also limits (self-reporting is highly subjective and directly testing the knowledge with limited available time would also be very problematic). Furthermore, we use the academic level mainly as a way to ensure some diversity among the users, rather than an exact proxy of knowledge of the domain.

was able to guide the users step by step through each required activity and each question.

Based of existing literature of the evaluation of search tools [153], multiple measures belonging to the two following categories was identified:

- subjective self-reported measures given by users, like quantitative answers on Likert scales or qualitative answers to open questions;
- objective measures, such as the log of the events from the user interface, the time to complete a task, and the words searched.

For the organization of the questions and the activities to be carried out by users during the test, the scheme proposed by Kelly [154] was followed. Kelly proposes to organize the questionnaires to evaluate interactive information retrieval systems (IIR) in five parts: demographic (e.g., gender, age), pre-task (e.g., prior knowledge of the system or topic), post-task (e.g., task satisfaction), post-system (e.g., the overall experience of interacting with an information system), and exit (e.g., cross-system comparisons of ease of use or preference). Below are the categories of questions used in this work.
User info:

- demographic (gender, age, level of education);
- pre-task (prior knowledge of relevant topics).

System evaluation (phases repeated for each of the four compared systems):

- task (the user is asked to navigate the system in order to retrieve a piece of information);
- post-task (quantitative evaluation of efficiency, effectiveness, and satisfaction to measure the usability).

Arca system in-depth evaluation:

- task (the user is asked to navigate the system in order to retrieve a piece of information);
- post-task (quantitative evaluation of efficiency, effectiveness, and satisfaction to measure the usability);
- post-system (the overall experience of interacting with Arca system).

To ensure the evaluation results' reliability, the following criteria was established:

- the whole process was executed through a self-administered web questionnaire, which ensured a level of distance between researcher and participant;
- users started directly with the comparative evaluation of the search systems without ever discussing Arca or its characteristics before;
- the four compared systems were presented in the same way and in a different order for each user group, asking them at the end of each navigation for feedback on their usefulness, satisfaction and ease of use.

To ensure the validity of each evaluation request's results, a clear objective on what to evaluate and with which metrics to do it best was established. For example, to measure usability, efficiency, effectiveness/usefulness and satisfaction with Likert scales with a rating of one to five was measured.

Key Factors. During the test, the proposed activities and subsequent questions were aimed at investigating the user interaction experience with the interface. In particular, the questions asked at the end of the activity tried to extrapolate an assessment of the key factors listed below.

- *Satisfaction. Namely, investigate how good is the system for the research objective, intended as the discovery and retrieval of information in a digital library.*
- *Effectiveness. How effective is the system in showing users the information.*
- *Support.* How much the system supports the user during the searches and exploration of a digital library.
- *Usefulness.* It directly impacts the usage of any system [163]. Therefore, usefulness can be considered a critical usability factor.
- *Learnability.* Analyze the way users adopt and get familiarized with the systems.

Use case. Inside Arca 112 books have been inserted concerning the history of Roman archaeology.

Users. Fifty-two people were selected for the test, including students and researchers from the domain of books contained within Arca.

Communication. All communications between the Arca team and the evaluating users took place by email.

The first email sent gave each user their login credentials and asked them to perform the three phases of the test:

- comparison of the four platforms proposed for searching books;

- in-depth evaluation of a single search platform (Arca);
- reply to a set of open questions on the whole process.

4.5.2 Results

The test, started in December 2020 and lasted for a month, is composed of four parts that were performed in this order:

1. the collection of personal data and self-assessment on knowledge background;
2. the comparison test of four book search platforms: Arca, Yewno, Lerma, Torrossa;
3. the evaluation test of Arca;
4. the test with open questions to express final evaluations.

This compilation order was chosen to respect the impartiality of judgment during the tests' execution: to not put Arca in a more favored position than the other search systems.

Regarding the number of participants who have completed each test part, there are:

- twenty-five users who completed all four parts;
- one user who has completed the first three test parts;
- four users who completed the first two parts.

On average, users completed the three parts of the test for more than an hour.

The event log, traced during the evaluation, revealed that most of the interactions with the interface. All interaction actions with the user interface are considered, excluding those that do not significantly affect the flow of the interaction (such as clicking on the buttons with the tutorials and consulting the search history). The 33.25% of user interactions concerned the search for terms (see Figure 4.10); 29.36% involved adding concepts to the whiteboard; 24.16% concerned the selection of elements; 6.38% concerned the elimination of concepts from the dashboard; 5.03% ("connections:loadLinks" and "connections:loadElements") concerned the exploration of the selected concepts; finally, 1.82% concerned modifying the searched keyword in the search bar.

The first part of the evaluation contains personal data and general information about the user's background on information visualization, knowledge graph and visual interfaces for querying and interacting with data. Figure 4.11 shows, with a Likert scale, that general knowledge on the required topics is very low.

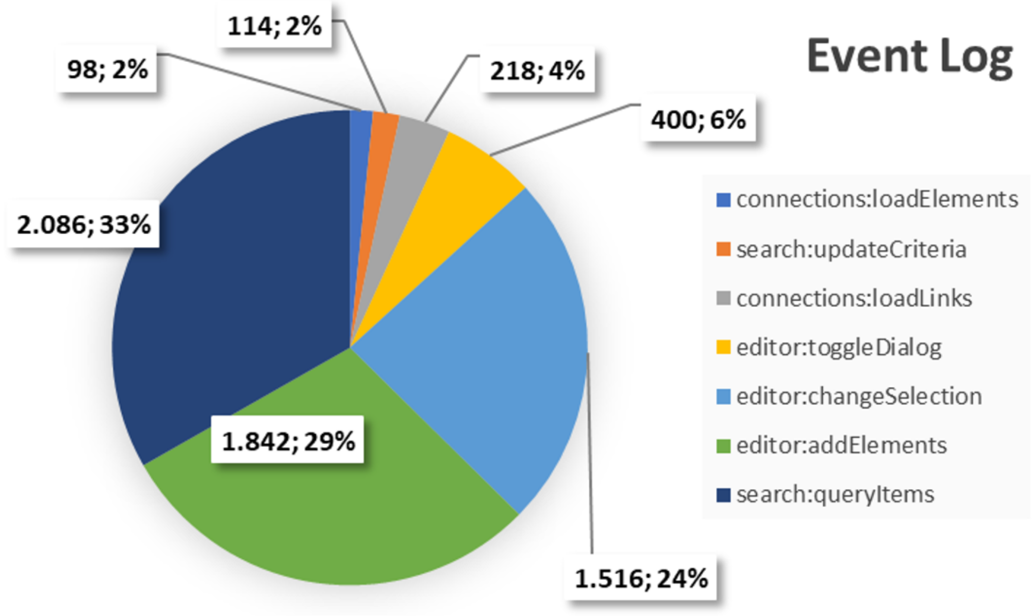


Figure 4.10. Percentage of event by type of action

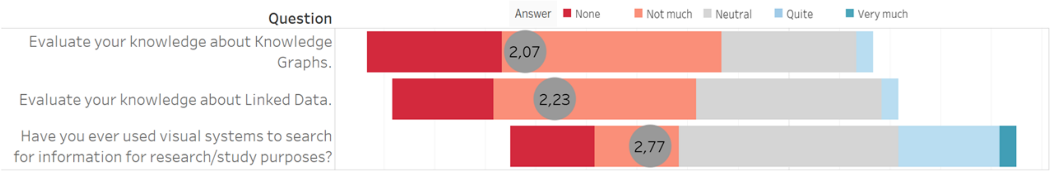


Figure 4.11. User background

Users who participated are mostly students and researchers in the humanities field. They are aged between twenty and forty-four, in particular 55% are between twenty and twenty-nine, 41% are between thirty and thirty-nine, 4% are over forty. Regarding the gender 49% are men and 51% women. Regarding education, 37% have a diploma, 40% have a degree, and 23% a Ph.D.

Comparative evaluation

After assigning the user the first search platform to test (randomly between Arca, Yewno, Lerma and Torrossa) and introducing him to its use, the task was to search for two books about two Roman hills. In Figure 4.12, the task one, associated with each search platform, indicate the search of the first book which talk about two Roman hills, while two indicates the search of the second book.

This task was created to encourage the user to follow multiple search paths without forcing him to a linear, and a sequential path to better evaluate the search

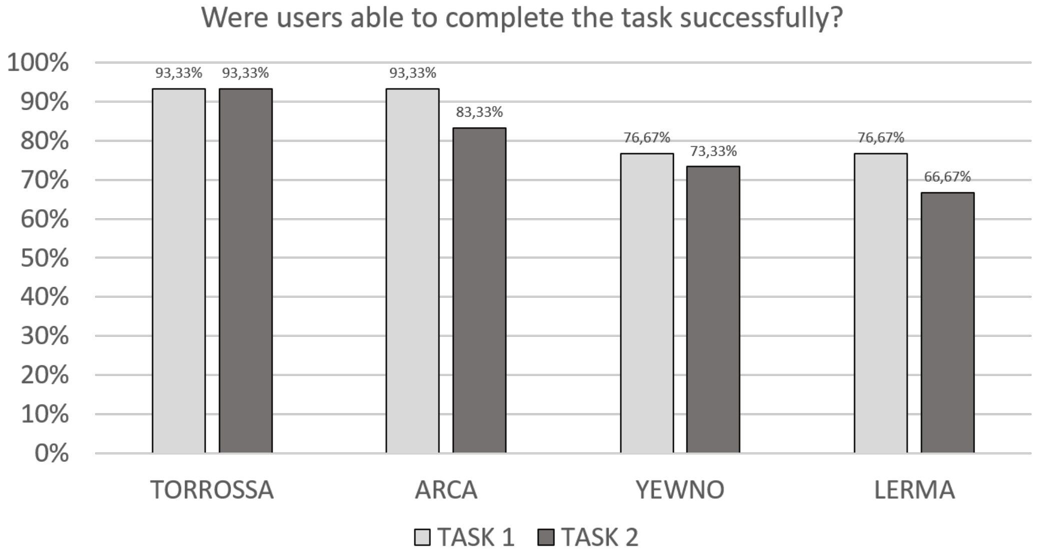


Figure 4.12. Completion of the comparative evaluation tasks

platform's usefulness, as explained in the research carried out by Liu et al. [155].

Not all users were able to complete the task, that is, to find the two required books. The positive results were checked to verify that the books' titles, indicated by users, actually contained two Roman hills. The return is that all users, who managed to complete the task, correctly indicated the books' titles.

After completing the task, the user was asked five quantitative questions with a Likert scale. The numerical score chosen for the likert scale is from one to five which represents the corresponding qualitative evaluation from "None" to "Very Much". Figures 4.13 and 4.14 show the distribution of the qualitative scores given to each research platform used: Lerma, Torrossa, Arca and Yewno. For the statistical analysis of the results, a one-way ANOVA was conducted to determine any statistically significant differences between the means of given scores to evaluate the key factors of satisfaction, effectiveness, support, usefulness and learnability for the four interfaces. Since the result of analysis of variance (F -value) is significant, it is necessary to use a Post-Hoc test [156], to identify which samples are different because the ANOVA test shows only that there is a difference between the means but does not indicate which means are different. The t-test is the selected Post-Hoc in this work. $p < 0.05$ was selected as the significance threshold.

Learnability

§The *learnability* factor was evaluated with the following post-task question: "How easy was it to complete the task?". The one-way ANOVA revealed that there was a statistically significant difference in mean value of ease in completing the research activity between at least four groups ($F(4.60, 1.09) = [4.20]$, $p = 0.01$).

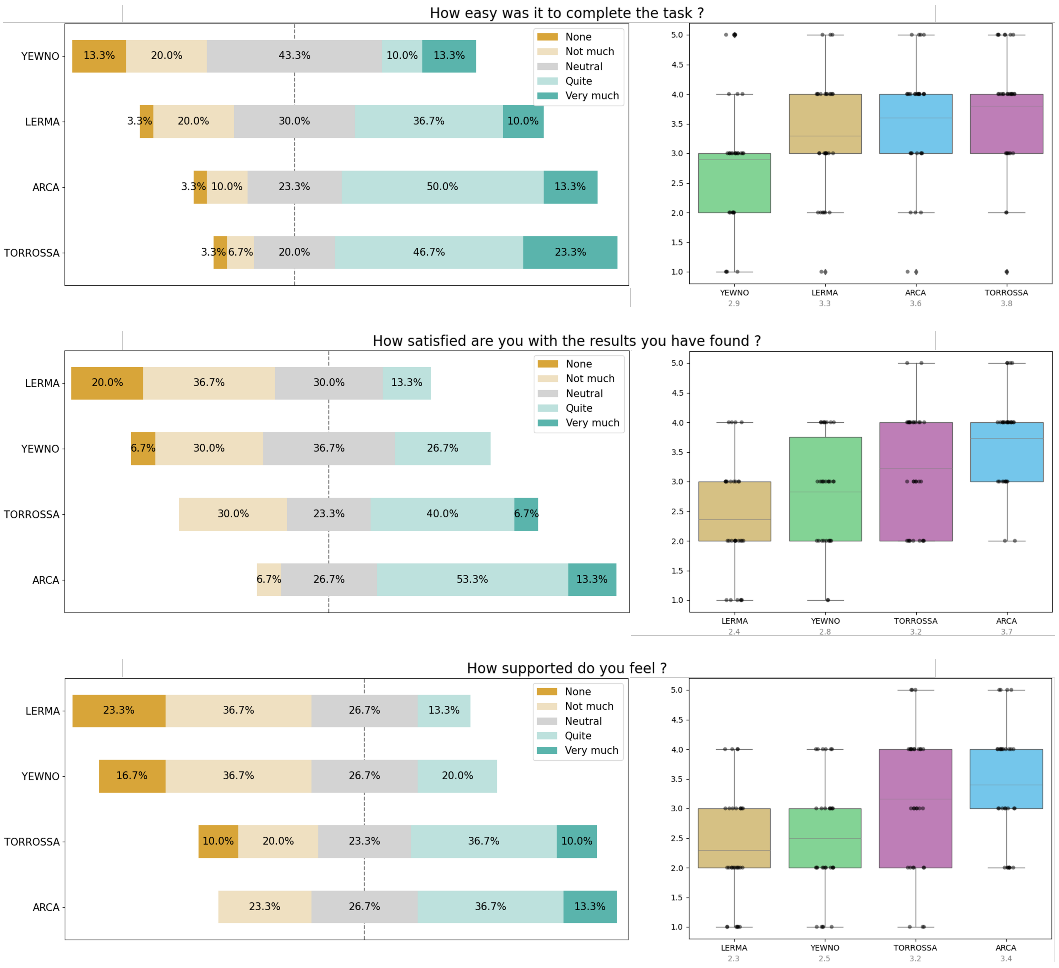


Figure 4.13. Comparative evaluation: distributions and averages - part 1

T-Test for multiple comparisons found that the mean value was significantly different between:

- Arca and Yewno (p = 0.01, statistics = 2.86)
- Lerma and Torrossa (p = 0.03, statistics = -2.29)
- Torrossa and Yewno (p = 8.32×10^{-4}, statistics = 3.73)

At the same time, there was no statistically significant difference between:

- Arca and Lerma (p = 0.17, statistics = 1.39)
- Arca and Torrossa (p = 0.33 , statistics = -1)
- Lerma and Yewno (p = 0.06 statistics = 1.93)

The results show no statistical significance between the ease of use of Arca compared to that of Lerma and Torrossa. This result underlines that although Arca is based on a graph information visualization system, which the user is less accustomed to, it is considered as easy as Lerma and Torrossa, systems based on schematic and tabular navigation. Furthermore, Arca is significantly better in learnability than Yewno, although both are based on displaying the results via a graph. However, Yewno does not allow the interactive exploration of the graph nodes, and maybe this is the reason for its lower ease of use compared to that of Arca.

Support

The *support* factor was evaluated with the following post-task question: “How supported do you feel?”. The one-way ANOVA revealed that there was a statistically significant difference in mean value of support in conducting the research activity between at least four groups (F(10.14, 0.83) = [12.20], p = 10^{-6}).

T-Test for multiple comparisons found that the mean value was significantly different between:

- Arca and Lerma (p = 1.54, statistics = 6.01)
- Arca and Torrossa (p = 0.02, statistics = 2.40)
- Arca and Yewno (p = 8.32×10^{-4}, statistics = 3.72)
- Lerma and Torrossa (p = 10^{-3}, statistics = -3.63)
- Lerma and Yewno (p = 0.02, statistics = -2.45)

At the same time, there was no statistically significant difference between:

- Torrossa and Yewno (p = 0.07 statistics = 1.88)

The tests show that users felt more supported by the Arca platform, with a significant statistical difference compared to the other three platforms tested. This could derive from the users' need (detected and satisfied) for more significant support because Arca has distinctive features compared to other tools for searching for information in a digital library. For example, the trace path allows users to find information common to several concepts; the snippets show the part of the text that deals with the concept explored; the graph exploration allows users to trace search paths and view connections directly on the dashboard.

Effectiveness

The *effectiveness* factor was evaluated with the following post-task question: "How satisfied are you with the results you have found?". The one-way ANOVA revealed that there was a statistically significant difference in mean value of the results shown during the research activity between at least four groups (F(8.28, 1.09) = [7.55], p = 1.17×10^{-4}).

T-Test for multiple comparisons found that the mean value was significantly different between:

- Arca and Lerma (p = 2.56×10^{-4}, statistics = 4.16)
- Arca and Yewno (p = 10^{-3}, statistics = 3.66)
- Lerma and Torrossa (p = 4.15×10^{-3}, statistics = -3.11)
- Torrossa and Yewno (p = 0.01 statistics = 2.94)

At the same time, there was no statistically significant difference between:

- Arca and Torrossa (p = 0.32, statistics = 1.02)
- Lerma and Yewno (p = 0.42, statistics = -0.81)

User Satisfaction

The *satisfaction* factor was evaluated with the following post-task question: "Is the information you viewed satisfactory for you?". The one-way ANOVA revealed that there was a statistically significant difference in mean value of the satisfaction perceived by the resulting information during the research process between at least four groups (F(7.83, 1.06) = [7.40], p = 1.41×10^{-4}).

T-Test for multiple comparisons found that the mean value was significantly different between:

- Arca and Lerma (p = 3.52, statistics = 4.88)
- Arca and Yewno (p = 2.61×10^{-4}, statistics = 4.16)
- Lerma and Torrossa (p = 2.94×10^{-3}, statistics = -3.25)

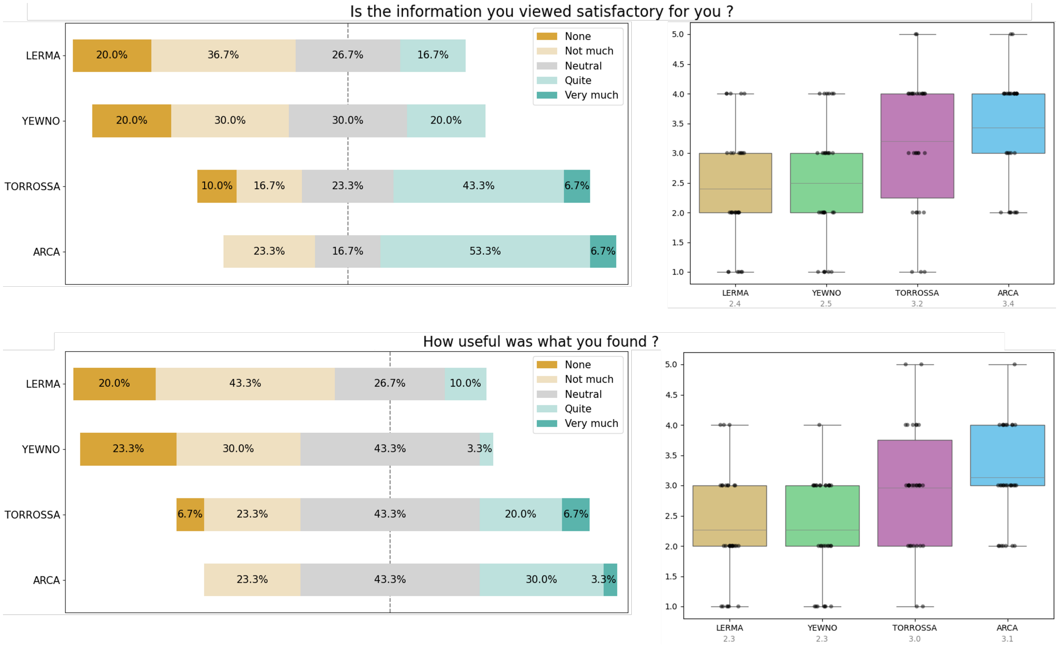


Figure 4.14. Comparative evaluation: distributions and averages - part 2

- Torrossa and Yewno (p = 0.01 statistics = 2.97)

At the same time, there was no statistically significant difference between:

- Arca and Torrossa (p = 0.27, statistics = 1.13)
- Lerma and Yewno (p = 0.68, statistics = -0.42)

As for the analysis of the satisfaction of the information shown (user satisfaction) and the results found (effectiveness), the user feels satisfied both with the use of Arca and Torrossa. On the contrary, the search results shown by Yewno and Lerma are considered less satisfactory by the user than the other two systems. It is possible to detect from the tests that although Yewno is based on the same technologies as Arca, this does not enhance it compared to Torrossa, which is still considered better in the information displayed and the search results. Maybe that Arca and Yewno, based on Knowledge Graph, semantic search and visualization of information on a graph, allow reaching more information and links than those Lerma and Torrossa, based on key-text search and visualization of tabular results. Despite the hypothesis, the results just commented support this for Arca, but not for Yewno. As already noted, this may be due to the explorability of the resources allowed by Arca.

Usefulness

The *usefulness* factor was evaluated with the following post-task question: "How useful was what you found?". The one-way ANOVA revealed that there was a statistically significant difference in mean value of perceived usefulness by the

resulting information during the research process between at least four groups (F(6.28, 0.81) = [7.73], $p = 9.4 \times 10^{-5}$).
T-Test for multiple comparisons found that the mean value was significantly different between:

- Arca and Lerma ($p = 2.47 \times 10^{-4}$, statistics = 4.18)
- Arca and Yewno ($p = 2.47 \times 10^{-4}$, statistics = 4.18)
- Lerma and Torrossa ($p = 2.9 \times 10^{-3}$, statistics = -3.25)
- Torrossa and Yewno ($p = 2.34 \times 10^{-3}$ statistics = 3.34)

At the same time, there was no statistically significant difference between:

- Arca and Torrossa (p = 0.39, statistics = 0.87)
- Lerma and Yewno (p = 1.0, statistics = 0.0)

The usefulness of the searches carried out with the Lerma, and Yewno tools are significantly lower than that carried out with the Torrossa and Arca tools. Torrossa and Arca are not significantly different. It is possible to deduce that in this case, the Knowledge Graph and semantic research have generated informative content as useful as Torrossa's manual metadata.

On average, it took users 26 minutes to complete the comparative evaluation. In particular, they sailed on average:

- 9.4 minutes Arca;
- 6.6 minutes Torrossa;
- 5.8 minutes Yewno;
- 4.3 minutes Lerma.

Referring to Figures 4.13 and 4.14, from the average of the scores assigned to each search platform, a list of the preference was derived. In order, users preferred the system (with a score from one [None] to five [Very much]):

- Arca with a score of 3.46;
- Torrossa with a score of 3.3;
- Yewno with a score of 2.6;
- Lerma with a score of 2.5.

The evaluation of Arca revealed that the system has good potential in:

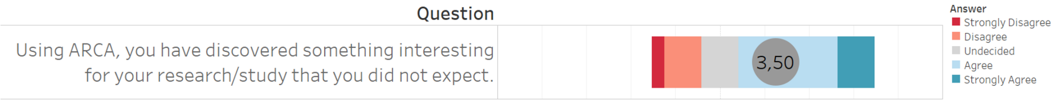


Figure 4.15. Serendipity evaluation

- in producing satisfactory results;
- in supporting the user in searches;
- in being useful to the user for his research;
- in producing satisfactory results for the user.

As for ease of use, users preferred the Torrossa search platform more.

Arca evaluation

For the evaluation of Arca, was set up research tasks with the aim of making users perform full navigation of Arca, exploring every component and every functionality offered by the system to make the most of its potential in order to reach the required research goal. Users perform guided navigation of the system to make them able to know all the functions and possibilities of search and navigation, and then was asked them for tasks and questions related to the research task just carried out.

Here are the three required tasks:

- Search and explore books about Rome in medieval times.
- Search and explore books about ancient Greek jewels.
- Research and explore books to deepen a topic, which is covered among the texts contained within Arca.

The tasks have been chosen to leave the user the freedom to explore the system according to their creativity. To evaluate this aspect and serendipity, was asked to users how favored they were at finding unexpected things (Figure 4.15).

Free navigation of the system resulted in neutral utility and neutral general satisfaction level. By analyzing the next section, will be explained the ideas for improving the system.

Final questions

The previous tasks have been chosen to leave the user the freedom to explore the system according to their creativity. To evaluate this aspect and serendipity, was asked to users how favored they were at finding unexpected things. Finally, three general questions were asked, shown below, along with an outline of user responses.

What are the most useful features of Arca?

- the possibility of observing the connections between books by finding arguments in common;
- the practicality that facilitates bibliographic research;
- the transversal approach to the topics;
- the visual search;
- semantic connections;
- wide-ranging exploration (bibliographic and conceptual);
- navigation of texts and concepts;
- the direct link to the catalogue of books for finding the resources of interest;
- the amplification of the results;
- the type of result display that facilitates complex searches;
- the simplicity of use;
- the graphic environment, although it can be improved, has good potential;
- the interdisciplinary research of contents;
- the ability to find books on more than one subject at a time, and the fact that the search is not limited to titles;
- the possibility to organize diagrams with all the connections from basic research to peripheral publications;
- being able to find texts with a common topic and above all with more topics in common.

What are Arca's weaknesses?

- difficult to use without having seen the tutorials;
- research does not always lead to what is sought;
- few results when searching for something specific;
- some links are non-existent;

- the lack of filters on searches.

Are there any features that could improve the attractiveness and usefulness of Arca?

- the introduction of components that allow more complex queries, such as the search for links in common to more than two resources;
- the inclusion of a more extensive catalogue of books to expand the number of links and information;
- investing in the graphics to make system navigation more comfortable and more intuitive;
- the increase of the tutorials and guides to allow the user to exploit the full potential of the system.

Chapter 5

Extensions of the proposed system

Contents

After the proposed system was positively evaluated (chapter 4), it was decided to proceed by developing some applications of potential interest for the domain of digital humanities and digital libraries as extensions of the primary system. In particular, the same interaction paradigm was applied to two different domains concerning the humanities (ancient symbols and ancient places). New functionality has been added downstream of the system that processes the images present in the digital library, recognizing the objects represented, which are inserted into the explorable knowledge domain. Also presented is the extension that allows domain experts to validate the quality of automatically extracted reports.

5.1 Same interaction paradigm / different domain

The availability of a tool such as the one proposed in this thesis would foster collaboration among the researchers in the area, and could attract curious [90],

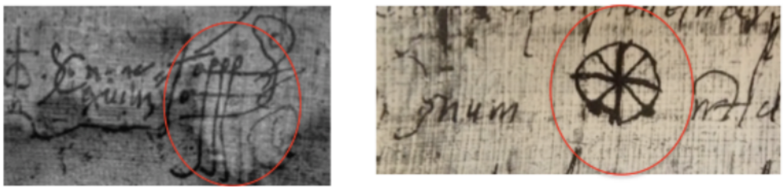


Figure 5.1. Examples of graphic symbols

casual, users by easing the diffusion of niche topics like those regarding ancient documentary texts.

5.1.1 A catalog of ancient manuscript

With the gradual introduction of subscription and the increasing use of papyrus the presence of graphic symbols became widespread in late Roman legal practice as it already was for some time in other expressions of the so-called “pragmatic literacy”; it continued in Post-Roman kingdoms as part of the same historical process of reception of the late antique documentary practice. A graphic symbol is meant as a graphic device composed by graphic signs (included alphabetical ones) drawn as a visual unit in a written text and representing something other or something more than a word of that text. The message it carries on is to be discovered, not only by definition, because there is no intrinsic prior relationship between the message-bearing graphic entity and the information it conveys, but also in historical terms: a huge quantity of graphic symbols of different forms (in appearance strange or common, unique or recurring, original or archetypal) was drawn by a huge quantity of persons, both literates of any degree (professional scribes, bureaucrats, economic élites) and illiterates, in different typologies of documents, with different functions and expectations, in different places, different social, cultural and economic contexts. Examples of graphic symbols are shown in Fig. 5.1: *Left)* graphic symbol in complex structure at the end of the subscription written by a Greek notary (Hermoupolis, Egypt, 561 AD). *Right)* symbol (Greek cross and diagonal cross crossing each other within a circle) drawn by an illiterate seller in his own hand on the sale contract (Ravenna, Italy, 572 AD).

For illiterates, for example, both in the Greek-Latin graphic and linguistic koinè of the late Roman state and in the Post-Roman kingdoms as long as Latin functioned as language of vertical communication, performing graphic symbols in their own hand certainly meant a way of taking an active part in the writing process: in other words, this phenomenon can be seen as the “other side” of the written world and considered as a matter of literacy in a wider sense. Not only for this reason the project Notae represents the first attempt to investigate this material from a novel perspective, as indicated in [157].

Notae system is the information system that let human experts involved in Notae to insert, update and query the research insights gathered during the project activities. In particular, by manual inspection of ancient documentary sources, researchers maintain a database of documents, physical supports (papyrus, wooden tablet, slate, parchment) containing texts and graphic symbols, if any, together with a wide range of related information. Part of this activity could be improved in future by the employment of automatic machine learning and image processing methods [158].

One of the foreseen outcomes for the project, is to discover geographical and historical implications of the employment of graphic symbols. During the implementation a particular attention was made on the Notae knowledge graph (KG).

Building the Notae KG is a way to:

- *(i)* introducing a common vocabulary for researchers in the area;
- *(ii)* sharing a common understanding of how concepts are related;
- *(iii)* enabling the reuse of domain knowledge;
- *(iv)* making domain assumptions explicit. In addition, to build a proper KG for the Notae project, a graphical user interface which allows researchers is also proposed:
  - *(i)* to explore the Notae KG;
  - *(ii)* to search and explore relations and connections between resources;
  - *(iii)* to make historical-geographical implications;
  - *(iv)* to discover new facts about the research field.

The Arca extension in the Notae context is a tool that can be employed in the broad field of digital humanities, conveying information not only from the Notae project but also from the vast set of related repositories and services available on the Internet. The challenge here is to enhance all of these resources, turning them into a semantically enriched ecosystem to ease information accessibility and knowledge discovery. Even though some of the available services allow to download data as Linked Open Data (LOD), none of them provides an integrated view of the research field nor allow to draw historical implications and explore data in a simple and intuitive manner. The availability of a tool such as the one proposed in this thesis would foster collaboration among the researchers in the area, and could attract curious, casual, users by easing the diffusion of niche topics like those regarding ancient documentary texts on which Notae project focuses.

The Notae knowledge graph

The Notae KG is built on top of the Notae Database, which consists of data manually inserted by domain experts as a result of their research activity in the context of the project. As the Notae Database is a relational database, an ETL process is executed daily to synchronize it with the KG.

The Notae KG makes use of multiple vocabularies, providing a set of classes and properties to describe the given domain. Vocabularies, expressed using the RDFS and OWL standards, make data integration easier by reducing diversity in describing things. The data model is defined incorporates our own vocabulary, described below, as well as the standard vocabularies listed below.

Schema.org[1] was born in 2011 by a collaboration of Google, Microsoft, Yahoo!, and Yandex to mark up website content with metadata about the website itself. Here the schema.org class "Place" is used.

DBpedia is one of the principal Linked Open Datasets[2], automatically extracted from multiple language versions of Wikipedia pages. The class "Language" is used from the DBpedia ontology.

FOAF[3] provides terms for describing people and organizations, documents associated with them, and social connections between people. Properties from FOAF are used to connect images of symbols to the Symbol class.

geo[4] provides terms for specifying geographical coordinates. Geo is used to define latitude and longitude of the instances of the class "Place".

An RDF schema is composed of properties and classes. Tables 5.1 and 5.2 list respectively classes and properties employed to define Notae KG.

The data contained in the knowledge graph are stored in a Blazegraph triple-storeand made available through a SPARQL endpoint[5]. For what concerns the visualization and the exploration of the Notae KG[6] (see Fig. 5.3), it is adapted the tool developed in the context of the ARCA project (see section:1.2.5).

Let's now see an example of how we can deduce historical implications from the data in the knowledge graph. Starting from two terms of interest (see Fig. 5.3), such as the symbol with ID 218 and Aphrodito, we can find all the entities that bind the two. In this case, the "Receipt" document. The same method can be applied to any class or instance of the graph. From this example, emerge how easy it is to find all documents that contain a symbol or to visually trace (by looking at the links

[1] see `http://schema.org/`

[2] see `http://dbpedia.org/ontology/` and `http://dbpedia.org/resource/`

[3] Friend Of A Friend - see `http://xmlns.com/foaf/0.1/`

[4] see `http://www.w3.org/2003/01/geo/`

[5] see `http://notae-system.diag.uniroma1.it:9999/blazegraph/#query`

[6] see `http://notae-system.diag.uniroma1.it:8888`. For reviewing purposes please use the following credentials, i.e., login: `IRCDL2021@notae.it` and password: `ircdl2021`.

Table 5.1. Notae KG Classes

name	description
FROM NOTAE NAMESPACE	
Document	A "document" is a written text generated for pragmatic purposes. Main focus of our research is the documentary records (such as contracts, lists, petitions, official and private letters, et cetera) dating from the 4th to the 8th centuries and written in Greek or Latin, on hard materials (such as slate fragments and wooden tablets) or on soft ones (papyrus, parchment).
Symbol	Graphic symbols are meant as graphic entities composed of graphic signs (included alphabetical ones) drawn as a visual unit within a written text but communicating something other or something more than a word of that text.
Category	"Category" defines the ideal geometric shape that a symbol designs or which can be inscribed in, for example, square, circle, and cross.
Shape	"Shape" is a way of defining a symbol based on its apparent structure. It can be simple, in the case of a single and easily recognizable symbol, or complex, in symbols with interconnected shapes or groups of symbols.
Element	An "element" is what a symbol is made of. It can be alphabetical, if the symbol is made of letters, tachygraphic, in case of stenography, and simple strokes for geometrical purposes.
Position	"Position" indicates the place in the doc. where a symbol is drawn: it specifies both the side and the exact line(s) of the doc. on which the symbol appears.
FROM OTHER NAMESPACES	
schema:Place	Indicates entities that have a somewhat fixed, physical extension.
dbo:Language	Represents the language of an object.

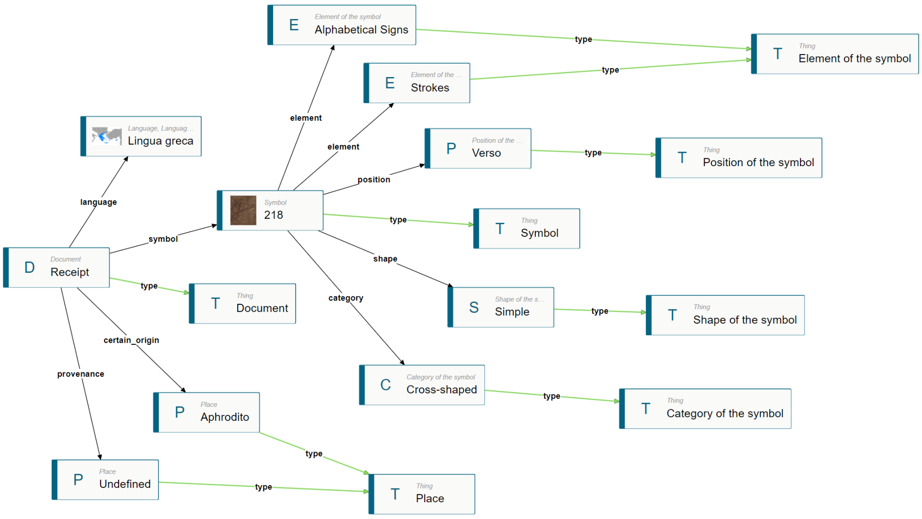


Figure 5.2. Notae KG exploration example.

Table 5.2. Notae KG Properties

name	description	domain	range	cardinality
FROM NOTAE NAMESPACES				
symbols	Connects a doc. to its symbols	Document	Symbol	multiple
provenance	Represents the provenance, i.e., the finding place of a document	Document	schema:Place	functional
shape	Represents the shape of a symbol	Symbol	Shape	functional
position	Represents the position of a symbol with respect to the document	Symbol	Position	functional
category	Is the category of a symbol	Symbol	Category	functional
element	Represents the graphical elements contained in a symbol	Symbol	Element	max(3)
certain_origin	Represents the certain origin of a document, i.e., its creation place	Document	schema:Place	functional
uncertain_origin	Is the uncertain origin of a doc.	Document	schema:Place	functional
undefined_origin	Is the undefined origin of a doc.	Document	schema:Place	functional
FROM OTHER NAMESPACES				
rdfs:label	Represents the human-readable name of a given resource	rdfs:Resource	rdfs:Literal	functional
dc:date	A point or period of time associated with the estimation of the document production	rdfs:Resource	rdfs:Literal	max(2)
dc:language	A language of the document	rdfs:Resource	dbo:Language	functional
dc:description	An account of the document	rdfs:Resource	rdfs:Resource	functional
foaf:depiction	Represents the image of a symbol	owl:Thing	dbo:Image	functional
geo:lat	The latitude of a SpatialThing	geo:SpatialThing	xmlns:basednear	functional
geo:lon	The longitude of a SpatialThing	geo:SpatialThing	xmlns:basednear	functional

that appear on the whiteboard) to the fact that a searched symbol is in a searched document.

5.1.2 A catalog of ancient maps

Another extension arises from the researchers' need to explore knowledge bases in a cartographic context. This is done by searching for links between the different topics related to a place or a key term, in such a way as to reveal unexpected connections during the exploration of contents and, thus, generating new ideas.

The Sciba platform, which can then be expanded on a national and international scale, will focus on the Lazio region (Italy), with the support of the “L’Erma di Bretschneider” publishing house, and will allow the visualization of the existing bibliography concerning some selected topographical elements. It will also be possible to expand the search to further related themes, automatically extracted from the system based on the contents and metadata of the integrated bibliographic material. The aim is to develop a search and comparative semantic analysis tool of great utility for research bodies, museums, municipalities and subjects interested in knowledge and management of the territory, and, at the same time, capable of generating an increase in volume sales with direct employment effects of the publishing chain.

The project:

- contributes to enriching and improving the technologies for the use of editorial material of historical and historical-artistic interest for the region;
- helps to facilitate the recognition of regional cultural heritage by the international public;
- contributes to the use of the Places of Culture of the Lazio Region;
- combines storytelling with academic research needs and the usability of cultural heritage;
- generates an increase in profitability for the industrial partners involved and concrete employment effects.

In summary, the Sciba platform offers a variety of results to ensure fruition, broaden access and encourage the discovery of cultural heritage. The project’s success rests on the network of academic institutions, places of culture and representatives of scientific publishing who find themselves on common objectives: the need for the application of new technologies for the use and enhancement of the Cultural Heritage. The possibility of implementing the contents and the availability of the same on a comprehensive platform ensures the increase in the visibility and impact of the project, which is configured as a pilot project at a national level and which will constitute a reference point for many researchers and institutions (academic companies, university press, independent publishing houses, commercial enterprises and all kinds of cultural institutions).

The Sciba platform, based on the Arca platform, arises from the researchers’ need to explore knowledge bases in a cartographic context. This is done by searching for links between the different topics related to a place or a key term, in such a way as to reveal unexpected connections during the exploration of contents and, thus, generating new ideas. These connections between concepts are visually represented by means of a graph that the user can arrange on her own, a solution already used in other projects, such as NOTAE (see section 5.1.1), a tool to investigate the graphic symbols in order to capture all the possible historical implications [73], which uses a similar visual-search system to navigate a Knowledge Graph.

The content search will be based on semantics and cartographic visualization. Unlike the classic search, where the search engine or application, based on keywords, proposes the texts, documents and metadata in which those words are present as a list of results, the semantic search has the aim of expanding the results and improve their accuracy by eliminating unnecessary results such as in the case of homographs or by including conceptually relevant results. To carry out semantic searches, it is necessary that metadata and texts are "annotated", that is, that unique identification codes (actually IRI, Internationalized Resource Identifier) are assigned to words contained in the texts, regardless of gender/number or, in the case of verbs, the inflected form and, in some cases, regardless of language. The actual search, then, will not be carried out on the single words but on the combination of semantic fields (the word, in all the genres and numbers in which it can be declined, synonyms, derivatives and the other words of the same semantic field).

Finally, by applying semantic search on a base cartography, it is possible to use a place name as a search key for a given theme. This is obtained by associating the cartography geometric and geographical information with the concept of "place" (i.e. a region or a point of interest), obtained from the reference texts with the use of an external AI. This AI will extract all the entities, represented by words from the texts, that will then linked by the platform to an external Knowledge Graph, specifically DBpedia, in such a way as to acquire the main information contained, included those concerning places. Geographical information of the places of interest are instead contained in the source maps files (i.e. GeoJSON file format[7]).

Once internal data (entities extracted from PDFs and maps, provided by a manager-role user) and the external information (DBpedia, linked by the platform) have been obtained, these two will be conveyed by the platform, with the use of a triple store, into a specifically created Knowledge Graph: the "Sciba knowledge graph" which can be explored in all its branches.

While the extraction of entities from the data, which is time-consuming, will have to be in batches, i.e., upstream of the whole process, every time the data sources (PDFs and maps) increase or undergo changes, the content search will be performed in real time and will have three different modes:

- **Location search:** search by area indicated on a map, displaying all concepts

[7] https://geojson.org/

related to it;

- **Entity search:** search for a word or a person within the database and extracts all the relationships that branch off from the main entity being searched;
- **Road map:** plots the links between the relations of two searched entities (origin and destination), allowing to 'navigate' in the topic of interest for subsequent approximations and reach other contents and documents on further potentially interesting topics.

The user of the Sciba platform, having tracked down all the contents of her interest, will be able to organize them according to her own needs and interests by querying the search engine, interacting with the expansion and multiplication of the information returned and exploring the virtual library and the graph.

5.2 Knowledge extraction from images

Another extension of the primary system involved the addition of the automatic extraction of the contents of the images present in a digital library. This extension has allowed the reuse of this information to generate new applications, for example, the semi-automatic creation of book trailers to support storytelling for digital libraries.

5.2.1 Semi-automatic video-trailer creation

StoryBook is a research project in the field of digital humanities, which proposes the knowledge extraction and management of information of a digital library to create semi-automatically video trailers of books. StoryBook has been originally conceived to meet the needs of "L'Erma di Bretschneider" publishing house that deals with topics related to ancient history and archaeology. From publishers' point of view, promotional trailers respond to a changing market with a high focus on digital and visual media. The goal of a digital presentation of a book is nevertheless broader than selling it and includes providing helpful information to the potential future reader. A publishing house wants to disclose the contents of its digital library not only to experts in the sector but also to interested people attracted by the contents of their books shown on the Web in the form of searching tools or advertising such as video trailers. Numerous researches show that a book trailer fosters the desire to learn and the level of motivation to read [83, 84, 85, 86]. StoryBook is a software tool to support the creation of book trailers by collecting and organizing relevant video content. The system users retain control on how to edit and compose the content. The proposed technique aims at semi-automatically building digital trailers that allow the viewers, generically interested in a specialized topic but not expert,

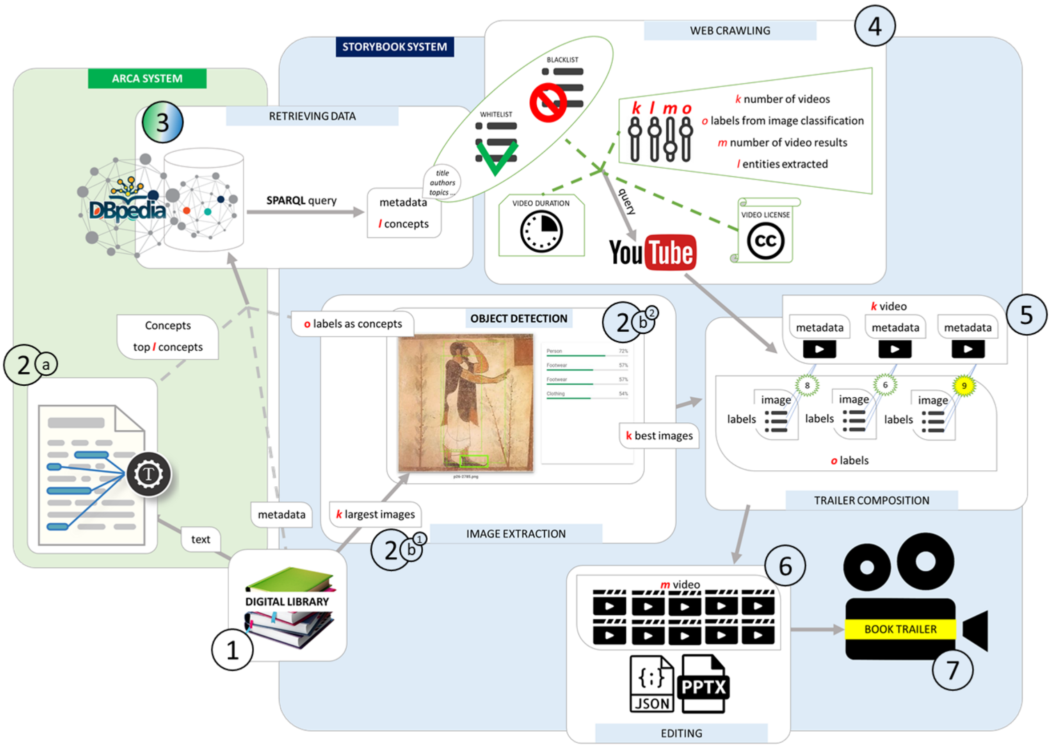


Figure 5.3. Storybook pipeline

to appreciate better the topic, both for their own cultural/professional enrichment and a possible purchase.

Querying a knowledge graph about the content of a digital library can semi-automatically generate a book's trailer based on this information, and can favor the diffusion of the contents of a digital library and enhance the cultural heritage contained in those digital libraries.

StoryBook potential is based on technologies such as:

- Linked Data to allow disambiguation of concepts and the connection of the information with the Web;
- Named entity recognition (NER) to identify people, cities, organizations and things in unstructured text like that of books;
- Knowledge graphs to organize information around concepts with their relations;
- Computer Vision to detect objects in the images of books allowing their searchability.

In Figure 5.3, there is the whole process that leads to the semi-automatic book trailer creation. The process starts from point 1 with a digital library (a collection of books in PDF format). Through the Arca platform, are extracted the concepts and the most relevant concepts of a book with Natural Language Processing (NLP) techniques. These concepts are linked with the DBpedia Knowledge Graph[8], and all this information are sent along with the relevant metadata to a linked data container (point 3 fig. 5.3).

In the system pipeline mages are extracted from books (point 2.b.1 fig. 5.3). The process is designed to work on books with meaningful visual content in images extracted from the PDF. We selected the "k" best images of the book. After trying different methods, the easiest one that yielded good results was sorting the images by their dimensions to detect the best images. It is observed that book publishers tend to let essential images take up more space on the page (thus being more significant in size once extracted).

In point 2.b.2 (fig. 5.3), there is object detection. In order to speed up the development, instead of implementing a custom image classification, is using an external service, Google Vision AI[9]. Up to "o" labels are retrieved for each image, along with their confidence level. The concepts and top "l" concepts associated with a book and its metadata coming from point 2.a (fig. 5.3), and the "o" labels, point 2.b.2 (fig. 5.3) associated with images of the book point 2.b.1 were sent as linked data to the linked data container (point 3 fig. 5.3).

At point 4 (fig. 5.3), now we can query different information from a book:

- all the concepts;
- the most relevant concepts;
- the metadata (like title, author, topics, etc.);
- the objects contained in the book's images.

Some parameters are introduced to allow users who are curators of a publishing house to filter the video search. The parameters are:

- the video duration;
- the type of video licence;
- a white list containing words relevant for curators;
- a black list containing the words that mustn't be in the search;

[8] https://dbpedia.org/

[9] https://cloud.google.com/

- the book’s metadata.

This information are used to Web crawl looking for relevant videos. After experimenting with different sources of video content, it was decided to focus on a single source, YouTube[10]. YouTube is currently the most extensive database of videos in the world. The query run on to YouTube filtered with all the elicited parameters.

The next step (point 5 fig. 5.3) of the process consists of filtering the “k” videos resulting from the previous step, and organising them in a draft of the final trailer. The algorithm assigns each video a multidimensional score (one dimension for each image). The score increases when there is a match between the image’s labels and the video’s metadata, such as the description. The confidence of the image-keywords association gives a score’s weight (as assigned during the classification step). Once each video has a score, the algorithm matches the highest score per single image, associates that video with the specific image, and discards all the others. Afterwards, the trailer is generated by interleaving the extracted images, and their correspondent retrieved videos.

This trailer draft is generated (point 6 fig. 5.3) for compatibility reasons as an annotated PowerPoint presentation, thus allowing the curator to manipulate the content as s/he think fitting before the actual video creation. Furthermore, all the data generated along the pipeline is packaged inside a JSON file that the curator can access to check the intermediates results and other details of the process.

Preliminary evaluation with a domain expert was carried on. Satisfactory results were obtained by choosing:

- “l” number of top concepts = “k” number of books’ images = “o” number of objects detected in one image = 10
- “m” number of resulted video’s query = 12

Furthermore, the domain expert gave some general feedback on the system paradigm. She identified several perceived strengths and potentialities of the system:

- the transformation of the images of books into information nodes;
- the control of the automatic creation process of a book trailer, through accessible configuration parameters;
- the free access and management of the information output generated by StoryBook.

[10] https://www.youtube.com/

She also expressed some concerns for the perceived weaknesses:

- the scarce availability of videos of niche topics on the Web;
- the prolonged extraction process duration for searches involving longer crawler queries.

5.3 Validation of the automatic knowledge extraction

The validation extension is an Arca component that, in addition to semantically extracting the concepts contained in a digital library (DL), allows the experts of the domain treated in the library to validate the information extracted. The automatic extraction thus represents a solid basis from which to intervene collaboratively by optimizing the quality of the associations between documents and their respective contents. Numerous researches [159, 160, 161] show that automatic extractions are insufficient to meet an acceptable quality level by communities of domain experts. At the same time, it would be too expensive in terms of time and cost to semantically annotate information manually. In this validation approach, the User Interface shows the semantic search result, the trace of the origin from where the information has been automatically extracted, and the history of user validation activities to facilitate the domain experts to validate the consistency of the relationships between automatically extracted concepts and a DL documents. For further information on the review of semantic annotations tools see section 3.6.

5.3.1 An approach to improve data quality

Improving the quality of data in a system that manages books, metadata, concepts, and images requires considering each of these entities and their associations.

The process of validation start with validating the associations between concepts and books by showing users in what terms the considered concept is described in the considered document, so a large percentage of errors due to completely automatic extractions can be eliminated in a short time.

This validation will be stored in the KG and the cloud, allowing immediate collaborative work. Validations could improve the search and display of results, and the system keeps track of the history of validations so the user can choose which information he prefers to view.

Figure 5.4 shows the semantic enrichment extraction and information validation processes pipeline: Users can validate or deny the relations between a corpus of texts and their relative concepts in the Knowledge extracted from a digital library. Once the artificial intelligence service has extracted the concepts from the text (named entity recognition) and from the images (object detection), the semantic enrichment

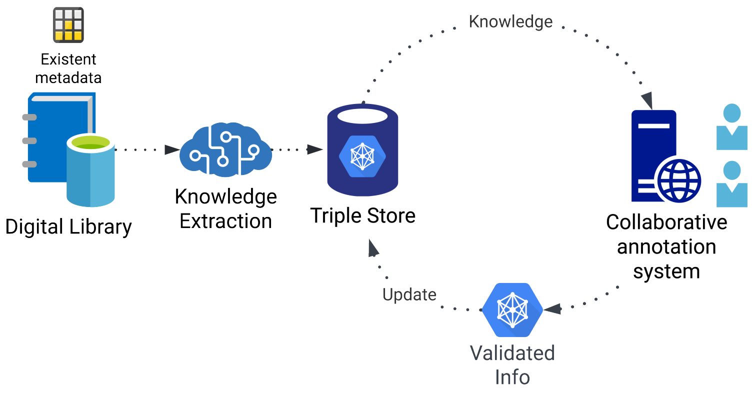


Figure 5.4. Collaborative validation pipeline.

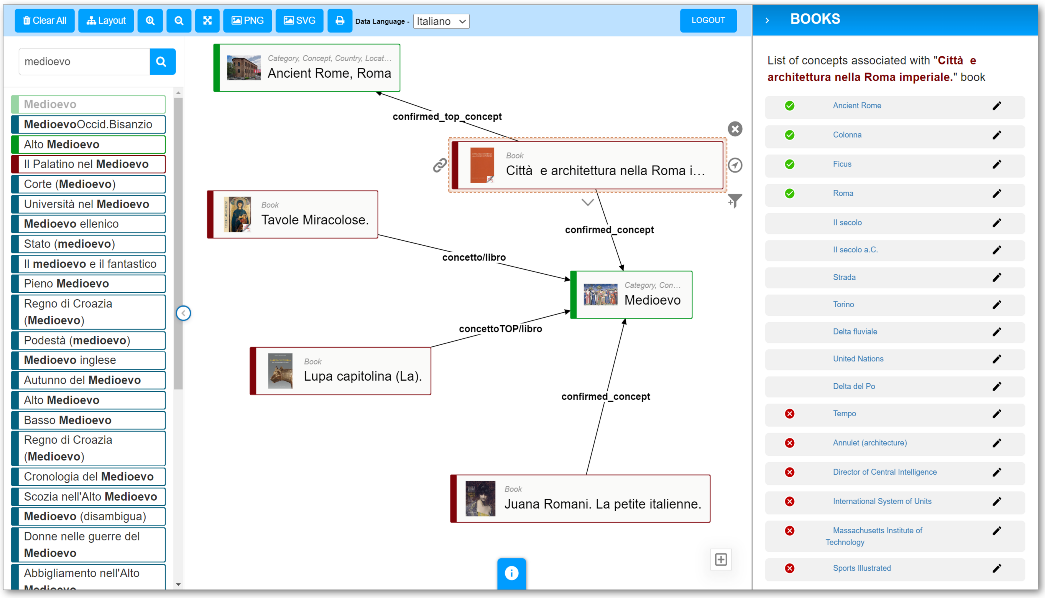
engine connects the identified concepts with the DBpedia knowledge base[11]. This mechanism allows visualizing the semantic associations between concepts. The validation process creates new associations between the ID of the user logged into the platform and the validated concept/book pair. These new associations are sent to the KG, which manages the validation versions from the different users of the system. With each new validation, the system updates the relationships, improving the results of the searches and the related visualization of the associations, establishing a process that aims to achieve the optimal quality for domain experts.

Figure 5.6 represents the tool's screenshot showing the books associated with the selected concept: "Middle Ages". Here the user can observe the book/concept relationships validated (green tick), those eliminated (red tick), and those automatically extracted (without ticks). With the edit symbol, the user can edit the relationship. Figure 5.5 represents the tool's screenshot showing the concepts associated with the selected book: "Città e architettura nella Roma imperiale".

Figure 5.6 represents the tool's screenshot showing the books associated with the selected concept: "Middle Ages". There is a tick associated with each book, which can be green for validated concept/book relationships and red for deleted relationships. Using the info button, the user can access the component (fig. 5.7) for editing the report and exploring the snippets, i.e. the phrases from the books containing the Middle Ages concept.

Figure 5.7 represents the tool's screenshot showing the snippets associated with the concept/book couple: "Middle Ages/Mura di Roma". Through this window, the user can establish whether the Rome concept is consistent with the theme dealt with in the book. To be guided in the decision, the user sees the book's text snippets (in this case, the sentences) from where the concept was automatically detected.

[11] `https://www.dbpedia.org/`



Medioevo
MedioevoOccid.Bisanzio
Alto Medioevo
Il Palatino nel Medioevo
Corte (Medioevo)
Università nel Medioevo
Medioevo ellenico
Stato (medioevo)
Il medioevo e il fantastico
Pieno Medioevo
Regno di Croazia (Medioevo)
Podestà (medioevo)
Medioevo inglese
Autunno del Medioevo
Alto Medioevo
Basso Medioevo
Regno di Croazia (Medioevo)
Cronologia del Medioevo
Scozia nell'Alto Medioevo
Medioevo (disambigua)
Donne nelle guerre del Medioevo
Abbigliamento nell'Alto Medioevo

› BOOKS

List of concepts associated with "**Città e architettura nella Roma imperiale.**" book

✓	Ancient Rome	✎
✓	Colonna	✎
✓	Ficus	✎
✓	Roma	✎
	Il secolo	✎
	Il secolo a.C.	✎
	Strada	✎
	Torino	✎
	Delta fluviale	✎
	United Nations	✎
	Delta del Po	✎
⊗	Tempo	✎
⊗	Annulet (architecture)	✎
⊗	Director of Central Intelligence	✎
⊗	International System of Units	✎
⊗	Massachusetts Institute of Technology	✎
⊗	Sports Illustrated	✎

Figure 5.5. Collaborative validation UI - part 1

Automatic semantic extraction includes the disambiguation of concepts, i.e. the choice of the best meaning based on the context of the sentence. For this reason, showing the text snippets to the user allows it to have all the information available to decide whether or not to validate the association.

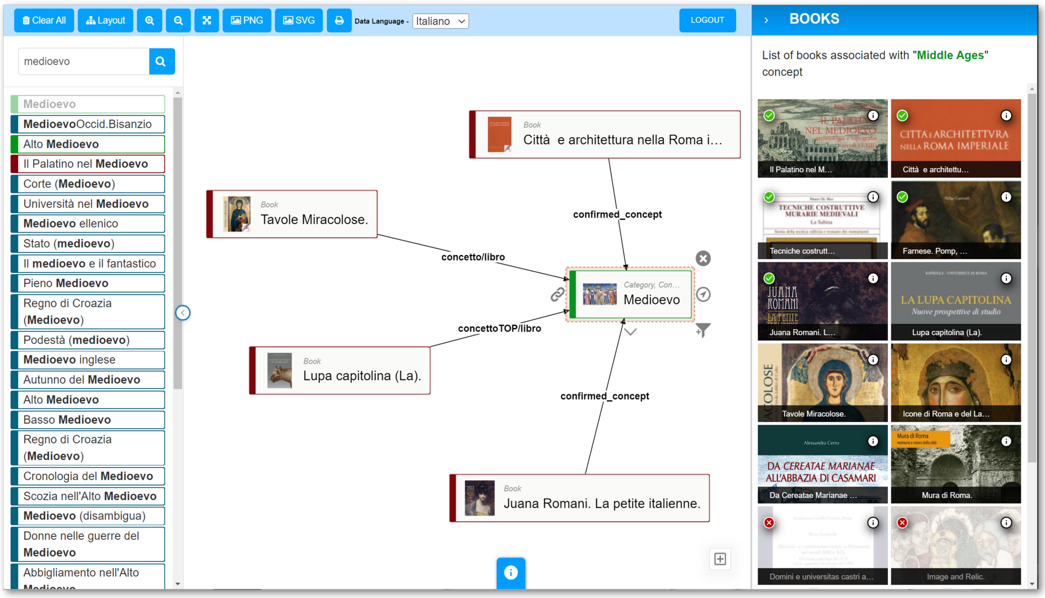
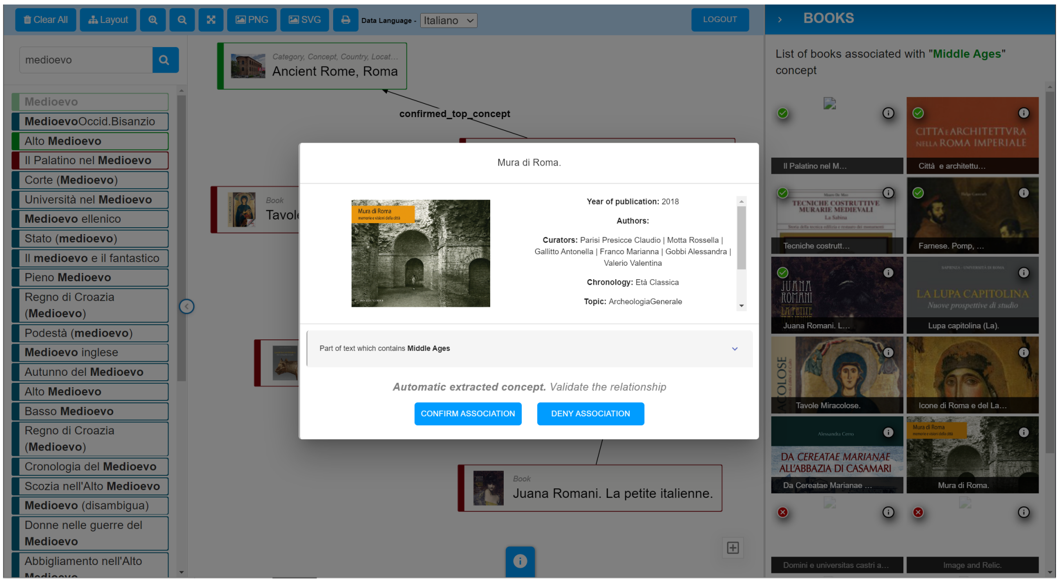


Figure 5.6. Collaborative validation UI - part 2



Mura di Roma.

Year of publication: 2018

Authors:

Curators: Parisi Presicce Claudio | Motta Rossella | Gallitto Antonella | Franco Marianna | Gobbi Alessandra | Valerio Valentina

Chronology: Età Classica

Topic: ArcheologiaGenerale

Part of text which contains **Middle Ages**

Automatic extracted concept. *Validate the relationship*

CONFIRM ASSOCIATION DENY ASSOCIATION

Figure 5.7. Collaborative validation UI - part 3

Chapter 6

Concluding remarks

Implementing an exploratory semantic search system requires various issues, including mapping text to semantic entities, detecting and cleaning inconsistencies in available linked open data, ranking algorithms for semantic data and heuristics for recommendations, and appropriate visualizations of complex semantic relationships.

This research has identified open challenges in the digital humanities research field related to:

- the user support in viewing, searching and exploring a digital library multimedia content
- the scalability of systems for extracting and exploring a book corpus;
- the difficulty of approaching these tools for users who are not experts in information technology;
- the quality of knowledge extraction;
- the difficulty in maintaining systems based on semantic technologies.

The research competence was demonstrated by studying user interaction with a designed and developed system based on automatic knowledge extraction, semantic technologies and visual search based on the serendipity effect to discover unexpected things. This system is proposed as a possible solution to the open problems identified in the research field of digital humanities.

6.1 Discussions and limitations

The system described in this thesis obtained a more than satisfactory performance in recommending relevant editorial products and received a high score in terms of usability; simplicity of use; user satisfaction with the results shown; consistency of the contents with the books domain of the publishing house; attractiveness of the system. Nonetheless some users identified as an issue the relatively small amount of information contained in the internal KG (built with the concepts of 112 books and the related metadata). It is expected that when the catalog of books is numerically more significant, the chance of discovering new information and connections while browsing the KG will increase.

Furthermore, based on initial observations, it has been seen that using the system at first glance can be difficult without viewing the video tutorials. In fact, concerning the comparative assessment, users, to solve the tasks, took longer to navigate on Arca than other tools. This fact may indicate a greater navigation complexity, but simultaneously, merged with positive feedback about usability, a greater exploratory interest. To reduce initial exploration difficulties, as a lighter alternative to video tutorials, a help component could be implemented to accompany the user in the first searches and thus make her independent in exploiting all the possibilities of exploration that the system offers.

Below the analysis of the test results will be discussed to support or re-evaluate the hypotheses elaborated in Section 4.2.3 and recalled in schema below.

Recalled hyphoteses 4.2.3

H1. Users will be able to effectively explore a text corpus through a KG-based user interface, which offers the following main functions: *a.* finding concepts through text search (among the ones pertinent to the specific domain), *b.* visually navigating the concepts and their relationships, and *c.* showing documents relevant to the selected concept.

H2. The method, given a corpus of texts in a specific domain, will benefit both users with little knowledge of the domain (by supporting semantically-relevant discovery) and domain experts (by enabling a topic-oriented visual organization of the documents).

H3. It is feasible to build a ready-to-use complete system, including both semantic enrichment pipeline and web-based front end, which is able, with only some configuration, to be applied to any specific corpus to enable the KG-based exploration.

The current findings appear to validate the H1 hypothesis. The users have evaluated multiple aspects of their experience more positively than other tools. Users stated that they had obtained useful results for their searches and could explore and search within a texts corpus. Crucially, the tool was rated as easy to use as text-based search tools, which employ a paradigm that is certainly far more familiar to the users. In the open questions section, many users have stated that they appreciate the opportunity to explore the resources and the possibility of observing the connections between concepts. So semantic technologies enable the user to achieve search objectives and produce and show more interconnected results than search engines that do not show semantic connections and relationships. Albeit the results relating to H1 are very promising, in a user study with limited available time is hard to evaluate advanced usage of semantic search for complex research scenarios and in-depth visual exploration of the knowledge graph. For that purpose, an in-use evaluation with a longer time span may be designed in future. The availability of public tools for semantic-based information retrieval will allow to collect data and user feedback on how they are used and in turn enable the design of better paradigms and tools.

Regarding hypothesis H2, as anticipated when describing the setup of the user study in Section 4.5.1, the academic level has been considered as a partial indicator of the level of knowledge of the field. In that sense, if the hypothesis holds we expect that usage of Arca to be satisfactory across different levels of academic qualification. As already described, the users rated positively the experience with Arca in respect to other tools. In none of the key factors the level of education correlates significatively with the perceived quality of the experience. The tracing of the logs shows that users with a higher education level (master's degree, doctorate) have dedicated on average 11.50% of total interactions to deepen concepts, while users with a lower level of education (diploma, three-year degree) devoted an average of 5.44%. This split of behavior hints at two different approaches to navigation:

- find specific resources;
- explore connections related to found resources.

It is possible that the approach to research of individual users determine this attitude in searching and exploring concepts.

Regarding hypothesis H3, the generality of the implemented prototype goes in the direction of proving the generalizability of the pipeline. Furthermore, part of the same pipeline is being applied to other different use cases, as described in 5.1. In conclusion, more work is required to fully evaluate the applicability of the tool in multiple contexts, but the results so far seem promising.

6.2 Future work

This thesis described an innovative system based on the visual semantic search and exploration of a text corpus. Through a knowledge graph-based navigation, the user can start from any relevant entity and reach other entities related to it, discovering in which books or articles each entity is present and evaluating which of these results are useful for their research. The user studies conducted so far confirm the amenability of the proposed system to domain experts which were able to perform non-trivial tasks of search and exploration, tasks that would be more cumbersome to execute with the search tools they are used to. Feedback gathered from users suggest that the proposed exploration mechanism tends to amplify the user experience by also offering opportunities for further study and discovery of sources, themes, and materials, which have the potential of enriching the research process with new ideas. In a comparative task-based evaluation with other tools for information retrieval on the same corpus, the users rated favorably multiple key factors of the experience with the system. Specifically, they rated it as easy to use as text-based search tools, notwithstanding the inherent complexity of the user interface due to richer functionality and novelty of the paradigm, and easier to use than another more static semantic-based visual tool presented to them.

More work is needed to better evaluate one of the stated hypotheses and aims of the tool, i.e. if the tool can be easily applied to other use cases and what is needed to improve its generality. Furthermore, as a potential future direction of research and development, the scope of the semantic enrichment process could be broadened to other document elements, such as images and captions, enriching KG exploration.

Bibliography

[1] Ahlberg, C., Shneiderman, B.: Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In: The Craft of Information Visualization, pp. 7–13. Elsevier (2003)

[2] Ahlberg, C., Williamson, C., Shneiderman, B.: Dynamic queries for information exploration: An implementation and evaluation. In: CHI '92, p. 619–626

[3] Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)

[4] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web Scientific American 284 (5), 34–43 (2001)

[5] Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: A survey of the state of the art. arXiv preprint arXiv:1601.08059 (2016)

[6] Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far Int. J. on Semantic Web and Information Systems 3 (5), 1–22 (2009)

[7] Bolina, M.: Yewno Discover. Nordic Journal of Information Literacy in Higher Education 11(1) (2019)

[8] Dadzie, A.S., Rowe, M.: Approaches to visualising linked data: A survey. Semantic Web 2(2), 89–124 (2011)

[9] Ehrlinger, L., Wöß, W.: Towards a Definition of Knowledge Graphs. Proc. of SEMANTiCS (Posters, Demos, SuCCESS) 48 (2016).

[10] Keim, D.A.: Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics 8(1), 1–8 (2002)

[11] Marie, N., Gandon, F.L.: Survey of linked data based exploration systems. In: IESD@ISWC (2014)

[12] Mouromtsev, D., Pavlov, D., Emelyanov, Y., Morozov, A., Razdyakonov, D., Galkin, M.: The simple web-based tool for visualization and sharing of semantic data and ontologies. In: ISWC 2015

[13] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1),3–26 (2007)

[14] Nisheva-Pavlova, M., Alexandrov, A.: GLOBDEF: A Framework for Dynamic Pipelines of Semantic Data Enrichment Tools. In: Proc. MTSR 2018, pp. 159–168

[15] Po, L., Bikakis, N., Desimoni, F., Papastefanatos, G.: Linked Data Visualization: Techniques, Tools, and Big Data, vol. 10. Morgan & Claypool Publishers, 2020

[16] Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics 36, 1–22 (2016)

[17] Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering 27(2), 443–460 (2014)

[18] Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: The craft of information visualization, pp. 364–371. Elsevier (2003)

[19] Singhal, A.: Introducing the Knowledge Graph: Things, not Strings, May 2012. https://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html [August, 2016]

[20] Zloof, M.M.: Query-by-Example: a data base language. IBM systems Journal 16(4), 324–343 (1977)

[21] Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T., Capanna, M.C., Fazio, C.D., Marcucci, R., Pender, E., Petriccione, F.M.: ARCA. Semantic exploration of a bookstore. Proceedings of the International Conference on Advanced Visual Interfaces. (2020).

[22] Gold, Matthew K., ed. Debates in the Digital Humanities. University of Minnesota Press, 2012. Minnesota Scholarship Online, 2015. doi: 10.5749/minnesota/9780816677948.001.0001.

[23] Berry D.M. (2012) Introduction: Understanding the Digital Humanities. Palgrave Macmillan, London. doi: 10.1057/9780230371934_1

[24] Drucker, J., The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship. Routledge . (2021)

[25] Hogan, A., et al.Knowledge Graphs. CoRR. https://arxiv.org/abs/2003.02320 (2020)

[26] Byrne, Kate (2007). Nested Named Entity Recognition in Historical Archive Text. International Conference on Semantic Computing (ICSC 2007), IEEE, https://doi.org/10.1109/icsc.2007.107

[27] Röder, M., Usbeck, R., Ngomo, A. N., (2018). GERBIL – Benchmarking Named Entity Recognition and Linking consistently. Semantic Web, 9(5), 605-625, ISSN 2210-4968, IOS Press, https://doi.org/10.3233/sw-170286

[28] Chaudhuri, Arindam, Mandaviya, Krupa, Badelia, Pratixa, Ghosh, Soumya K. (2016). Optical Character Recognition Systems. Optical Character Recognition Systems for Different Languages with Soft Computing, 9-41, ISSN 1434-9922, Springer International Publishing, https://doi.org/10.1007/978-3-319-50252-6_2

[29] Huynh, Vinh-Nam, Hamdi, Ahmed, Doucet, Antoine (2020). When to Use OCR Post-correction for Named Entity Recognition?. Digital Libraries at Times of Massive Societal Transition, 33-42, ISSN 0302-9743, Springer International Publishing, https://doi.org/10.1007/978-3-030-64452-9_3

[30] Gramatica, Ruggero, Pickering, Ruth (2017). Start-up story Yewno: an AI-driven path to a knowledge-based future. Insights the UKSG journal, 30(2), 107-111, ISSN 2048-7754, Ubiquity Press, Ltd., https://doi.org/10.1629/uksg.369

[31] Patrício, Helena Simões, Cordeiro, Maria Inês, Ramos, Pedro Nogueira (2020). From the web of bibliographic data to the web of bibliographic meaning: structuring, interlinking and validating ontologies on the semantic web. International Journal of Metadata, Semantics and Ontologies, 14(2), 124, ISSN 1744-2621, Inderscience Publishers, https://doi.org/10.1504/ijmso.2020.108318

[32] (2014). Bibliographic Information Organization in the Semantic Web. Online Information Review, 38(6), 832-833, ISSN 1468-4527, Emerald, https://doi.org/10.1108/oir-08-2014-0180

[33] Bizer, Christian, Heath, Tom, Berners-Lee, Tim (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22, ISSN 1552-6283, IGI Global, https://doi.org/10.4018/jswis.2009081901

[34] Miller, Eric (2005). An Introduction to the Resource Description Framework. Bulletin of the American Society for Information Science and Technology, 25(1), 15-19, ISSN 0095-4403, Wiley, https://doi.org/10.1002/bult.105

[35] Rico, M., Vila-Suero, Daniel, Botezan, Iuliana, Gómez-Pérez, A. (2019). Evaluating the impact of semantic technologies on bibliographic systems: A user-centred and comparative approach., https://doi.org/10.1016/J.WEBSEM.2019.03.001

[36] Horrocks, I. (2000). Benchmark Analysis with FaCT. In: Dyckhoff, R. (eds) Automated Reasoning with Analytic Tableaux and Related Methods. TABLEAUX 2000. Lecture Notes in Computer Science, vol 1847. Springer, Berlin, Heidelberg., https://doi.org/10.1007/10722086_6

[37] Halpin, H., Hayes, P.J., McCusker, .P., McGuinness, D.L., Thompson, H.S. (2010). When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: , et al. The Semantic Web – ISWC 2010. ISWC 2010. Lecture Notes in Computer Science, vol 6496. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17746-0_20

[38] Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.L. (2010). SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. In: , et al. The Semantic Web – ISWC 2010. ISWC 2010. Lecture Notes in Computer Science, vol 6496. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-17746-0_10

[39] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010, volume 628 of CEUR Workshop Proceedings. CEUR-WS.org, 2010.

[40] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. J. Web Sem., 14:14–44, 2012.

[41] A. Soulet, A. Giacometti, B. Markhoff, and F. M. Suchanek. Representativeness of Knowledge Bases with the Generalized Benford's Law. In The Semantic Web – ISWC 2018 – 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I, pages 374–390, 2018.

[42] S. Issa, P. Paris, and F. Hamdi. Assessing the Completeness Evolution of DBpedia: A Case Study. In S. de Cesare and U. Frank, editors, Advances in Conceptual Modeling – ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6–9, 2017, Proceedings, volume 10651 of LNCS, pages 238–247. Springer, 2017.

[43] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for Linked Data: A Survey. Semantic Web, 7(1):63–93, 2016.

[44] M. Nentwig, M. Hartung, A. N. Ngomo, and E. Rahm. A survey of current Link Discovery frameworks. Semantic Web, 8(3):419–436, 2017.

[45] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009). Discovering and Maintaining Links on the Web of Data. In: , et al. The Semantic Web - ISWC 2009. ISWC 2009. Lecture Notes in Computer Science, vol 5823. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04930-9_41

[46] A. N. Ngomo and S. Auer. LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In T. Walsh, editor, IJCAI 2011, Pro 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16–22, 2011, pages 2312–2317. IJCAI/AAAI, 2011.

[47] V. Lopez, V. S. Uren, M. Sabou, and E. Motta. Is Question Answering fit for the Semantic Web?: A survey. Semantic Web, 2(2):125–155, 2011.

[48] C. Unger, A. Freitas, and P. Cimiano. An Introduction to Question Answering over Linked Data. In M. Koubarakis, G. B. Stamou, G. Stoilos, I. Horrocks, P. G. Kolaitis, G. Lausen, and G. Weikum, editors, Reasoning Web. Reasoning on the Web in the Big Data Era – 10th International Summer School 2014, Athens, Greece, September 8–13, 2014. Proceedings, volume 8714 of LNCS, pages 100–140. Springer, 2014.

[49] E. Oren, R. Delbru, and S. Decker. Extending Faceted Navigation for RDF Data. In I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, editors, The Semantic Web – ISWC 2006, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5–9, 2006, Proceedings, volume 4273 of LNCS, pages 559–572. Springer, 2006.

[50] M. Arenas, B. C. Grau, E. Kharlamov, S. Marciuska, and D. Zheleznyakov. Faceted search over RDF-based knowledge graphs. J. Web Semant., 37-38:55–74, 2016.

[51] J. Moreno-Vega and A. Hogan. GraFa: Scalable Faceted Browsing for RDF Graphs. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee, and E. Simperl, editors, The Semantic Web – ISWC 2018 – 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I, volume 11136 of LNCS, pages 301–317. Springer, 2018.

[52] Butuc, M. G. (2009). Semantically enriching content using opencalais. Editia, 9, 77-88.

[53] E. L. McCarthy, B. P. Vandervalk, and M. Wilkinson. SPARQL Assist language-neutral query composer. BMC Bioinform., 13(S-1):S2, 2012.

[54] L. Rietveld and R. Hoekstra. The YASGUI family of SPARQL clients. Semantic Web, 8(3):373–383, 2017.

[55] Malyshev, S., Krötzsch, M., González, L., Gonsior, J., Bielefeldt, A. (2018). Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph. SEMWEB.

[56] Vargas, H., Buil-Aranda, C., Hogan, A., López, C. (2019). RDF Explorer: A Visual SPARQL Query Builder. In: , et al. The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science(), vol 11778. Springer, Cham. https://doi.org/10.1007/978-3-030-30793-6_37

[57] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. Commun. ACM, 57(10):78–85, 2014.

[58] S. Battle, D. Wood, J. Leigh, and L. Ruth. The Callimachus Project: RDFa as a Web Template Language. In J. F. Sequeda, A. Harth, and O. Hartig, editors, Proceedings of the Third International Workshop on Consuming Linked Data, COLD 2012, Boston, MA, USA, November 12, 2012, volume 905 of CEUR Workshop Proceedings. CEUR-WS.org, 2012.

[59] M. E. Gutiérrez, N. Mihindukulasooriya, and R. García-Castro. LDP4j: A framework for the development of interoperable read-write Linked Data applications. In R. Verborgh and E. Mannens, editors, Proceedings of the ISWC Developers Workshop 2014, co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014, volume 1268 of CEUR Workshop Proceedings, pages 61–66. CEUR-WS.org, 2014.

[60] Loseto, G., Ieva, S., Gramegna, F., Ruta, M., Scioscia, F., Di Sciascio, E. (2016). Linked Data (in Low-Resource) Platforms: A Mapping for Constrained Application Protocol. In: , et al. The Semantic Web – ISWC 2016. ISWC 2016. Lecture Notes in Computer Science(), vol 9982. Springer, Cham. https://doi.org/10.1007/978-3-319-46547-0_14

[61] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A. N. Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, and A. Zimmermann. Knowledge Graphs. CoRR, abs/2003.02320, 2020.

[62] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry scale Knowledge Graphs: Lessons and Challenges. ACM Queue, 17(2):20, 2019.

[63] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia – A large-scale,multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2):167–195, 2015.

[64] H. Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3):489–508, 2017.

[65] Carriero, V.A. et al. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. In: , et al. The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science(), vol 11779. Springer, Cham. https://doi.org/10.1007/978-3-030-30796-7_3

[66] Szekely, P. et al. (2015). Building and Using a Knowledge Graph to Combat Human Trafficking. In: , et al. The Semantic Web - ISWC 2015. ISWC 2015. Lecture Notes in Computer Science, vol 9367. Springer, Cham. https://doi.org/10.1007/978-3-319-25010-6_12

[67] Kärle, E., Şimşek, U., Panasiuk, O., Fensel, D. (2018). Building an Ecosystem for the Tyrolean Tourism Knowledge Graph. In: Pautasso, C., Sánchez-Figueroa, F., Systä, K., Murillo Rodríguez, J. (eds) Current Trends in Web Engineering. ICWE 2018. Lecture Notes in Computer Science(), vol 11153. Springer, Cham. https://doi.org/10.1007/978-3-030-03056-8_25

[68] Dee Andy Michel. 1994. What is used during cognitive processing in information retrieval and library searching? eleven sources of search information. J. Am. Soc. Inf. Sci. 45, 7 (Aug. 1994), 498–514.

[69] Bernasconi, E., Ceriani, M., Mecella, M., Morvillo, A. Automatic Knowledge Extraction from a Digital Library and Collaborative Validation. International Conference on Theory and Practice of Digital Libraries Springer, Cham. pp. 480–484 Padua (September 2022)

[70] Bernasconi, E., Ceriani, M., Mecella, M., De Luzi, F., Sapio, F. StoryBook. Automatic generation of book trailers. International Conference on Advanced Visual Interfaces Association for Computing Machinery. pp. 1-3 Frascati (June 2022)

[71] Bernasconi, E., Ceriani, M., De Luzi, F., Sapio, F., Mecella, M. Storybook: a tool for the semi-automatic creation of book trailers.Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

[72] Bernasconi, E., Ceriani, M., De Luzi, F., Di Fazio, C., Marcucci, R., Mecella, M., Sapio, F. StoryBook-A Storytelling-based Platform for Digital Book Stores.Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

[73] Bernasconi, E., Boccuzzi, M., Catarci, T., Ceriani, M., Ghignoli, A., Leotta, F., Ziran, Z. Exploring the Historical Context of Graphic Symbols: the NOTAE Knowledge Graph and its Visual Interface. 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 147-154 Padua (February 2021)

[74] Bernasconi, E., Boccuccia, P., Fabbri, M., Francescangeli, A., Marcucci, R., Mecella, M., Morvillo, A., Tondi, E. SCIBA-A Prototype of the Computerized Cartographic System of an Archaeological Bibliography.Workshop. International Conference on Research Challenges in Information Science. Ceur-ws. Barcelona (May 2022)

[75] Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T. Design, realization and user evaluation of the ARCA system for exploring a digital library.Journal. International Journal on Digital Libraries Springer. (2022)

[76] Bernasconi, E., Ceriani, M., Mecella, M. Exploring a Text Corpus via a Knowledge Graph. 17th Italian Research Conference on Digital Libraries Ceur-ws. pp. 91-102 Padua (February 2021)

[77] Ceriani, M., Bernasconi, E., Mecella, M. A streamlined pipeline to enable the semantic exploration of a bookstore. Italian Research Conference on Digital Libraries Springer, Cham. pp. 75-81 Bari (January 2020)

[78] Bernasconi, E., Ceriani, M., Mecella, M. Academic Research Creativity Archive (ARCA). International Conference on Research Challenges in Information Science. Springer, Cham. pp. 713-714 (May 2021)

[79] A. Ghignoli, The notae project: a research between est and west, late antiquity and early middle ages, Comparative Oriental Manuscript Studies Bullettin 5/1 (2019) 27–39.doi:https://doi.org/10.25592/uhhfdm.185.

[80] Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T., Capanna, C., Di Fazio C., Marcucci, R., Pender, E., Petriccione, F.: ARCA. Semantic exploration of a bookstore. (AVI '20). Association for Computing Machinery, New York, NY, USA, Article 78, pp. 1–3. (2020).

[81] Ceriani, M., Bernasconi, E., Mecella, M.: A Streamlined Pipeline to Enable the Semantic Exploration of a Bookstore. (IRCDL 2020). Springer International Publishing, Cham, pp. 75–81. (2020).

[82] Bernasconi, E., Ceriani, M., Mecella, M.: Exploring a Text Corpus via a Knowledge Graph. (IRCDL 2021). CEUR Workshop Proceedings, pp. 91–102. (2021).

[83] Nikonova, Nadezhda Ilinichna, Zalutskaya, Svetlana Yrievna (2021). Modern communication technologies in education: book trailer. Revista Tempos e Espaços em Educação, 14(33), ISSN 2358-1425, Revista Tempos e Espacos em Educacao.

[84] Chepukov, K.Yu. (2021). Expressive means of painting as a means of self-expression of younger school children. ISSN 2072-0432, International Centre Art and Education.

[85] Jiménez-Marín, Gloria, Zambrano, Rodrigo Elías (2020). The Book Trailer as a Publishing House Promotional Tool. Advances in Business Strategy and Competitive Advantage, 147-160, ISSN 2327-3429, IGI Global.

[86] D. Luchev, D. Paneva-Marinova, M. Dimova (2019) Digital Storytelling and digital book trailer applications for educational purposes in Bulgaria, INTED2019 Proceedings, pp. 529-534.

[87] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics 36 (2016), 1–22.

[88] Maria Nisheva Pavlova and Asen Alexandrov. 2020. Extending the GLOBDEF framework with support for semantic enhancement of various data formats. International Journal of Metadata, Semantics and Ontologies 14, 2 (2020), 158–158.

[89] Esko Ikkala, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2021. Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web (2021), 1–16.

[90] Mecella, M., Leotta, F., Marrella, A., Palucci, F., Seri, C., Catarci, T.: Encouraging persons to visit cultural sites through mini-games. EAI Endorsed Trans. Serious Games 4(14), e3 (2018)

[91] Bodenhamer, David J., John Corrigan and Trevor M. Harris. "The Spatial Humanities: GIS and the Future of Humanities Scholarship." (2010).

[92] Alvarado, Rafael C. (2012). The Digital Humanities Situation. Debates in the Digital Humanities, 50-55, University of Minnesota Press, https://doi.org/10.5749/minnesota/9780816677948.003.0005

[93] Mariana Neves and Jurica Ševa. An extensive review of tools for manual annotation of documents. Briefings in Bioinformatics, 22(1):146–163, 12 2019.

[94] Amit Kumar and Marc Spaniol. Annotag: Concise content annotation via lod tags derived from entity-level analytics. In Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen, editors, Linking Theory and Practice of Digital Libraries, pages 175–180. Springer International Publishing, 2021.

[95] Cejuela, J. M., McQuilton, P., Ponting, L., Marygold, S. J., Stefancsik, R., Millburn, G. H., FlyBase Consortium. (2014). tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. Database, 2014.

[96] Andrea Loreggia, Simone Mosco, and Alberto Zerbinati. Sentag: A web-based tool for semantic annotation of textual documents. In Gerd Berget, Mark Michael Hall, Daniel Brenn, and Sanna Kumpulainen, editors, ThirtySixth AAAI Conference on Artificial Intelligence. AAAI Press, June 2022.

[97] Gleyzes, M., Perret, L., Kubik, P.: Pleiades system architecture and main performances. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp. 537–542 (2012)

[98] Broux, Y., Depauw, M.: Developing onomastic gazetteers and prosopographies for the ancient world through named entity recognition and graph visualization: Some examples from trismegistos people. In: SocInfo Workshops (2014)

[99] Depauw, M., Gheldof, T.: Trismegistos: An interdisciplinary platform for ancient world texts and related information. pp. 40–52 (07 2014)

[100] saksen, L., Simon, R., Barker, E.T.E., de Soto Can?amares, P.: Pelagios and the emerging graph of ancient world data. In: WebSci '14: Proceedings of the 2014 ACM conference on Web science. pp. 197–201. ACM (June 2014)

[101] Segel, J. Heer, Narrative visualization: Telling stories with data, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1139–1148. doi:10.1109/tvcg.2010.179

[102] C. Tong, R. Roberts, R. Borgo, S. Walton, R. Laramee, K. Wegba, A. Lu, Y. Wang, H. Qu, Q. Luo, X. Ma, Storytelling and visualization: An extended survey, Information 9 (2018) 65–65. URL: http://dx.doi.org/10.3390/info9030065. doi:10.3390/info9030065

[103] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, D. Zhang, Datashot: Automatic generation of fact sheets from tabular data, IEEE transactions on visualization and computer graphics (2019) 895–905

[104] A. Satyanarayan, J. Heer, Authoring narrative visualizations with ellipsis, Computer Graphics Forum 33 (2014) 361–370. URL: http://dx.doi.org/10.1111/cgf.12392.

[105] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, H. Qu, Infonice, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018) doi:10.1145/3173574.3173909

[106] D. Ren, M. Brehmer, B. Lee, T. Hollerer, E. K. Choe, Chartaccent: Annotation for data-driven storytelling, 2017 IEEE Pacific Visualization Symposium (PacificVis) (2017). URL: http:// dx.doi.org/10.1109/pacificvis.2017.8031599

[107] Q. Wang, Z. Li, S. Fu, W. Cui, H. Qu, Narvis: Authoring narrative slideshows for introducing data visualization designs, IEEE Transactions on Visualization and Computer Graphics 25 (2019) 779–788.

[108] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, P. Irani, Authoring data-driven videos with dataclips, IEEE Transactions on Visualization and Computer Graphics 23 (2017) 501–510.doi:10.1109/tvcg. 2016.2598647

[109] S. Bocconi, F. Nack, L. Hardman, Vox populi: a tool for automatically generating video documentaries, in: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, 2005, pp. 292–294.

[110] I. Teixeira, P. Viana, M. Andrade, P. Carvalho, L. Vilaça, J. Pinto, T. Costa, S. Rapanakis, P. Jonker, Creating automatic storytelling videos from still images: a semantically-aware approach, 2021.

[111] T. Gao, J. R. Hullman, E. Adar, B. Hecht, N. Diakopoulos, Newsviews: an automated pipeline for creating custom geovisualizations for news, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2014, pp. 3005–3014

[112] Simon, Rainer et al. “Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2.” Journal of Map Geography Libraries 13 (2017): 111 - 132.

[113] P.N. Mendes, M. Jakob, A. García-Silva and C. Bizer, DBpedia spotlight: Shedding light on the Web of Documents, in: International Conference on Semantic Systems (I-Semantics), C. Ghidini, A.-C. Ngonga Ngomo, S.N. Lindstaedt and T. Pellegrini, eds, ACM, 2011, pp. 1–8. doi:10.1145/2063518.2063519.

[114] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal,M. Spaniol, B. Taneva, S. Thater and G. Weikum, Robust disambiguation of named entities in text, in: Empirical Methods in Natural Language Processing (EMNLP), ACL, 2011,pp. 782–792.

[115] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov and A. Kirilov, KIM – a semantic platform for information extraction and retrieval, Natural Language Engineering 10(3–4) (2004),375–392. doi:10.1017/S135132490400347X.

[116] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R.V. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J.A. Tomlin and J.Y. Zien, Semtag and seeker: Bootstrapping the Semantic Web via automated semantic annotation, in: World Wide Web Conference (WWW), G. Hencsey, B. White, Y.R. Chen, L. Kovács and S. Lawrence, eds, ACM, 2003, pp. 178–186. doi:10.1145/775152.775178.

[117] A.G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, P. Ciancarini, Aemoo:exploring knowledge on the web, in: WebSci, 2013.

[118] I. Santana-Pérez, Graphless: Using statistical analysis and heuristics for visualizing large datasets, in: VOILA@ISWC, in: CEUR Workshop Proceedings, vol. 2187, CEUR-WS.org, 2018, pp. 1–12.

[119] N. Bikakis, J. Liagouris, M. Kromida, G. Papastefanatos, T.K. Sellis, Towards scalable visual exploration of very large RDF graphs, in: F. Gandon, C. Guéret, S. Villata, J.G. Breslin, C. Faron-Zucker, A. Zimmermann (Eds.), The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events PortoroŽ, Slovenia, May 31–June 4, 2015, Revised Selected Papers, in: Lecture Notes in Computer Science, vol. 9341, Springer, 2015, pp. 9–13.

[120] N. Bikakis, J. Liagouris, M. Krommyda, G.Papastefanatos, T.K. Sellis,GraphVizdb: A scalable platform for interactive large graph visualization,in: 32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16–20, 2016, IEEE Computer Society, 2016, pp.1342–1345.

[121] L. Po, High-level visualization over big linked data, in: M. van Erp, M.Atre, V. López, K. Srinivas, C. Fortuna (Eds.), Proceedings of the ISWC 2018 Posters Demonstrations, Industry and Blue Sky Ideas Tracks Co-Located with 17th International Semantic Web Conference, ISWC 2018, Monterey,USA, October 8th - to - 12th, 2018, in: CEUR Workshop Proceedings, vol.2180, CEUR-WS.org, 2018.

[122] M. Weise, S. Lohmann, F. Haag, LD-VOWL: Extracting and visualizing schema information for linked data endpoints, in: Proceedings of the 2nd International Workshop on Visualization and Interaction for Ontologies and Linked Data, VOILA 2016, in: CEUR-WS, vol. 1704, CEUR-WS.org,2016, pp. 120–127.

[123] D.V. Camarda, S. Mazzini, A. Antonuccio, Lodlive, exploring the web of data, in: V. Presutti, H.S. Pinto (Eds.), I-SEMANTICS 2012 - 8th International Conference on Semantic Systems, I-SEMANTICS '12, Graz,Austria, September 5–7, 2012, ACM, 2012, pp. 197–200.

[124] A. Micsik, Z. Tóth, S. Turbucz, LODmilla: Shared visualization of linked open data, in: Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P.Manghi (Eds.), Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops, Springer International Publishing, Heidelberg, 2014, pp. 89–100.

[125] A. Micsik, S. Turbucz, A. Györök, LODmilla: a linked data browser for all, in: S. Harald, F. Agata, L. Jens, H. Sebastian (Eds.),Posters Demos@SEMANTiCS 2014, CEUR-WS.org, 2014, pp. 31–34.

[126] P. Bellini, P. Nesi, A. Venturi, Linked open graph: Browsing multiple SPARQL entry points to build your own LOD views, J. Vis. Lang. Comput. 25 (6) (2014) 703–716, Distributed Multimedia Systems DMS2014 Part I.

[127] F. Haag, S. Lohmann, S. Siek, T. Ertl, QueryVOWL: Visual composition of SPARQL queries, in: Proceedings of ESWC 2015 Satellite Events, in: LNCS, vol. 9341, Springer, 2015, pp. 62–66.

[128] F. Haag, S. Lohmann, S. Siek, T. Ertl, QueryVOWL: A visual query notation for linked data, in: Proceedings of ESWC 2015 Satellite Events, in: LNCS, vol. 9341, Springer, 2015, pp. 387–402.

[129] S. Lohmann, V. Link, E. Marbach, S. Negru, WebVOWL: Web-based Visualization of Ontologies, in: EKAW, 2014

[130] R. Chawuthai, H. Takeda, RDF Graph Visualization by Interpreting Linked Data as Knowledge, in: JIST, 2015.

[131] G. Troullinou, H. Kondylakis, E. Daskalaki, D. Plexousakis, RDF digest: Efficient summarization of RDF/s KBs, in: ESWC, in: Lecture Notes in Computer Science, vol. 9088, Springer, 2015, pp. 119–134.

[132] P. Heim, S. Lohmann, T. Stegemann, Interactive relationship discovery via the semantic web, in: L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, T. Tudorache (Eds.), The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30–June 3, 2010, Proceedings, Part I, in: Lecture Notes in Computer Science, vol. 6088, Springer, 2010,pp. 303–317.

[133] F. Viola, L. Roffia, F. Antoniazzi, A. D'Elia, C. Aguzzi, T. Salmon Cinotti, Interactive 3D exploration of RDF graphs through semantic planes, Future Internet 10 (8) (2018).

[134] G. Tartari, A. Hogan, WiSP: Weighted shortest paths for RDF graphs, in: VOILA@ISWC, in: CEUR Workshop Proceedings, vol. 2187, CEUR-WS.org,2018, pp. 37–52.

[135] Yanan Zhang, Gong Cheng, and Yuzhong Qu. Towards exploratory relationship search: A clustering-based approach. In Semantic Technology—Joint International Conference, JIST,pp. 277–293, 2013. DOI: 10.1007/978-3-319-06826-8_21 57, 58

[136] C.B. Neto, K. Müller, M. Brümmer, D. Kontokostas, S. Hellmann, LODVader: An interface to LOD visualization, analyticsand discovery in real-time, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B.Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web,WWW 2016, Montreal, Canada, April 11–15, 2016, Companion Volume,ACM, 2016, pp. 163–166.

[137] Tuukka Hastrup, Richard Cyganiak, and Uldis Bojars. Browsing linked data with fenfire. In International World Wide Web Conference (WWW), 2008. 57, 58

[138] Marie, N., Gandon, F., Ribiere, M., Rodio, F.: Discovery hub: on-the-fly linked data exploratory search. In: Proceedings of the 9th Int. Conf. on Semantic Systems. pp.17–24. ACM (2013)

[139] C. Anutariya and R. Dangol, "VizLOD: Schema Extraction And Visualization Of Linked Open Data," 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2018, pp. 1-6, doi: 10.1109/JCSSE.2018.8457325.

[140] Jesse C. J. van Dam, Jasper J. Koehorst, Peter J. Schaap, Vitor Martins dos Santos, and María Suárez-Diez. RDF2Graph a tool to recover, understand and validate the

ontology of an RDF resource. Journal of Biomedical Semantics, 6:39, 2015. DOI: 10.1186/s13326-015-0038-9 58,63

[141] Florian Haag, Steffen Lohmann, and Thomas Ertl. SparqlFilterFlow: SPARQL query composition for everyone. In Extended Semantic Web Conference (ESWC), pp. 362–367, 2014c. DOI:10.1007/978-3-319-11955-7_49 58, 64

[142] Chakravarthy, M. A., Ciravegna, P. F., Lanfranchi, M. V. (2006). AKTiveMedia: Cross-media document annotation and enrichment.

[143] Hogue, A., Karger, D. (2005, May). Thresher: automating the unwrapping of semantic content from the world wide web. In Proceedings of the 14th international conference on World Wide Web (pp. 86-95).

[144] Giannopoulos, G., Bikakis, N., Dalamagas, T., Sellis, T. (2010, May). GoNTogle: a tool for semantic annotation and search. In Extended Semantic Web Conference (pp. 376-380). Springer, Berlin, Heidelberg.

[145] Gangemi, A. (2013, May). A comparison of knowledge extraction tools for the semantic web. In Extended semantic web conference (pp. 351-366). Springer, Berlin, Heidelberg.

[146] Aidan Hogan. 2020. Linked Data. The Web of Data (2020), 515–625. https://doi.org/10.1007/978-3-030-51580-5_8

[147] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. Now Publishers. https://doi.org/10.1561/9781680837292

[148] Prud'hommeaux, E., Carothers, G., Beckett, D., Berners-Lee, T. (2014). RDF 1.1 Turtle: terse RDF triple language. W3C recommendation, 25, 2008-2014.

[149] RDF Schema. (2020). In Semantic Web for the Working Ontologist. ACM. https://doi.org/10.1145/3382097.3382106

[150] Parsia, B., Patel-Schneider, P., Motik, B. (2012). OWL 2 web ontology language structural specification and functional-style syntax. W3C, W3C Recommendation.

[151] Yu, L. (2014). SPARQL: Querying the Semantic Web. In A Developer's Guide to the Semantic Web (pp. 265–353). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43796-4_6

[152] Bernasconi, E., Ceriani, M., Mecella, M. (2021). Arca - Evaluation (1.0) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.4679175

[153] O'Brien, H. L., McCay-Peet, L. (2017, March). Asking" good" questions: Questionnaire design and analysis in interactive information retrieval research. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (pp. 27-36).

[154] Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1–2), 1-224.

[155] Liu, J., Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. Journal of the Association for Information Science and Technology, 66(1), 58-81.

[156] Stemler, S. E., Tsai, J., Osborne, J. (2008). Best practices in quantitative methods.

[157] Ghignoli, A.: The NOTAE Project: a Research between Est and West, Late Antiquity and Early Middle Ages. Comparative Oriental Manuscript Studies Bullettin 5/1, 27–41 (2019)

[158] Boccuzzi, M., Catarci, T., Deodati, L., Fantoli, A., Ghignoli, A., Leotta, F., Mecella, M., Monte, A., Sietis, N.: Identifying, classifying and searching graphic symbols in the notae system. In: Italian Research Conference on Digital Libraries. pp.111–122. Springer (2020)

[159] Beall, J.: Metadata and data quality problems in the digital library. Journal of Digital Information 6(3) (2005),https://journals.tdl.org/jodi/index.php/jodi/article/view/65

[160] Candela, G., Escobar, P., Carrasco, R.C., Marco-Such, M.: Evaluating the quality of linked open data in digital libraries. Journal of Information Science 48, 21–43 (2022). https://doi.org/10.1177/0165551520930951

[161] Hallo, M., Luj an-Mora, S., Mat e, A., Trujillo, J.: Current state of linked data in digital libraries. Journal of Information Science 42, 117–127 (2016)

[162] Guha, R. V., Brickley, D., Macbeth, S. (2016). Schema. org: evolution of structured data on the web. Communications of the ACM, 59(2), 44-51.

[163] G Tsakonas and C Papatheodorou. 2006. Analysing and evaluating usefulness and usability in electronic information services. (2006). https://doi.org/10.1177/0165551506065934