KB-Driven Information Extraction to Enable Distant Reading of Museum Collections

Giovanni A. Cignoni, Progetto HMR, giovanni.cignoni@di.unipi.it

Enrico Meloni, Progetto HMR, enrico-meloni@hotmail.it

Short abstract (same as the one submitted in ConfTool)

Distant reading is based on the possibility to count data, to graph and to map them, to visualize the relations which are inside the data, to enable new reading perspectives by the use of technologies. By contrast, *close reading* accesses information by staying at a very fine detail level. Sometimes, we are bound to close reading because technologies, while available, are not applied.

Collections preserved in the museums represent a relevant part of our cultural heritage. Cataloguing is the traditional way used to document a collection. Traditionally, catalogues are lists of records: for each piece in the collection there is a record which holds information on the object. Thus, a catalogue is just a table, a very simple and flat data structure. Each collection has its own catalogue and, even if standards exists, they are often not applied.

The actual availability and usability of information in collections is impaired by the need of accessing each catalogue individually and browse it as a list of records. This can be considered a situation of mandatory close reading.

The paper discusses how it will be possible to enable distant reading of collections. Our proposed solution is based on a *knowledge base* and on *KB-driven information extraction*.

As a case study, we refer to a particular domain of cultural heritage: *history of information science and technology* (HIST). The information corpus about HIST is particularly suitable to distant reading methods. Moreover, being information technology pervasive of everyone life we need new ways of telling HIST: distant reading may help to engage people in the discovery of HIST and in the understanding of the science behind today informatics and in the comprehension of cultural and social phenomena generated by use and habit of these technologies.

1. Introduction

Today technologies make new reading approaches viable. *Distant reading* is based on the possibility to count data, to graph and to map them, to visualize the relations which are inside the data, thus enabling new reading perspectives (Moretti 2005). By contrast, *close reading*, is the way to access information by staying at a very fine detail level. There are cases in which we are bound to close reading because the technologies are available but not applied.

A relevant part of our cultural heritage is made of the collections preserved in museums. Cataloguing is the traditional way used to keep and to convey documentation about a collection. Catalogues are lists of records: for each piece in the collection there is a record which keeps information about the physicality of the object (the size, the materials, the state of conservation...), its origin (the author, the place where it was manufactured), its day by day management (its place in the museum exhibits o in the deposits, if it is on loan...).

As a data structure, a traditional museum catalogue is just a table, that is a very simple and flat data structure – we could say primitive. Moreover, each collection has its own catalogue and often, even if standards exist, it adopts its own schema for the data fields in the table.

The actual availability and usability of information about cultural heritage in collections is impaired by the need of accessing the catalogues individually and then by browsing them as lists of records. We could say that we are in a situation of *mandatory close reading*.

The paper discusses how it will be possible to enable distant reading of collections. We will refer to a particular domain of scientific and technological heritage: the *history of information science and technology* (HIST). Apart of being our field of historical interest, it is an interesting case study. First of all, the information corpus about HIST is particularly suitable to be investigated with distant reading methods. Moreover, being information technology pervasive of everyone life, there is interest – as well as need – for new ways of telling HIST: distant reading may help to engage people in the discovery of the many stories of HIST, in the understanding of the science behind today informatics and in the comprehension of the cultural and social phenomena generated by use and habit of these technologies.

Section 2 of the paper presents our specific domain and defines the requirements for enabling distant reading of HIST starting from the museum collections. Section 3 describes the solution we are proposing which is based on a *knowledge base* (KB) and on *KB-driven information extraction* (KB-DIE). Section 4 is devoted to discuss related approaches and methods and to analyse the feasibility of the implementation of our solution.

2. Distant reading of collections, the HIST case study

Information technologies are born in the Fifties and have quickly flourished in the following years. While they can be considered recent, they pervasively affect our everyday life, thus they are a proper cultural heritage of humanity – and not only as a scientific and technological matter.

As a natural consequence of the impact of informatics on the society, there is a growing curiosity about HIST: protagonists such Alan Turing or Steve Jobs have become pop icons celebrated in popular movies (Tyldum 2014, Stern 2013, Boyle 2015). At the same time, the interest in the conservation of HIST relics is raised. Important collections belong to museums generally devoted to science and technology, such as, among the best known, the *Science Museum*[1] of London, the *Deutsches Museum*[2] in Munchen, the *Conservatoire National des Arts et Métiers*[3] in Paris, the *Museum of Science and Industry*[4] in Manchester. There are also museums specifically dedicated to HIST like the *Computer History Museum*[5] in Mountain View, the *Heinz Nixdorf Museumforum*[6] in Paderborn, or *The National Museum*

[1] http://www.sciencemuseum.org.uk/ (last accessed November 13, 2016).

[2] http://www.deutsches-museum.de/en/exhibitions/communication/computers/ (last accessed November 13, 2016).

[3] http://www.cnam.fr/ (last accessed November 13, 2016).

[4] http://msimanchester.org.uk/ (last accessed November 13, 2016).

[5] http://www.computerhistory.org/ (last accessed November 13, 2016).

[6] https://www.hnf.de/en/museum.html (last accessed November 13, 2016).

of Computing[7] at Bletchley Park. In addition to museums there are collections belonging to enthusiasts, often organized in *retro-computing* clubs and associations. People involved are usually very cooperative, often also very competent on specific topics. While not open to the public on a regular basis, such collections represent an important contribute to research and preservation about HIST. In the following, collections have to be intended in a wide and participative way.

2.1. The starting point, from catalogues to KB

The idea of passing from many different, flat catalogues of collections to a unique shared KB has already been presented, as a general idea and as a feasible international project (Cignoni and Cossu 2016). A very basic prototype, namely *CHKB*[8], has been developed as part of several students' theses at the University of Pisa to investigate how a simple KB structure can be implemented and how users could populate it through a web interface. A KB has several advantages. It keeps the complexity of relations among the facts of the HIST domain. Being unique, the KB collects all the knowledge and makes it surfable at different levels, both for experts and researchers and for the general public. The KB is also an authoritative source: all the facts pass through a peer review revision process by an open group of experts which cooperates to assure a high level of reliability of the KB content. For the details, we refer to the cited works, here we focus on how such a KB can allow distant reading of HIST.

2.2. Distant reading of HIST by enriching the KB contents

Distant reading may greatly support the work of HIST scholars. The domain is characterized by many complex relations. From the perspective of relics conservation, the KB explicits the relation among the pieces in the collections and the product they are instances of. For example, from the KB it is possible to build the map of all the preserved *Apple][* telling which of them are on exhibition, which of them are in working conditions, which of them are periodically demonstrated.

From a more historical perspective, the Apple][was a product of *Apple*, it has a predecessor (the *Apple-1*) and a successor (the *Apple ///*), it has a designer (*Steve Wozniak*). Even more interesting in the KB are the relations concerning the software or hardware components. For instance, the Apple][uses the *MOS 6502* microprocessor like many other devices of the time which belong to different categories: *personal computers* (like the Apple ones as well as the *Commodore PET* series, or the less known *Ohio Scientific Challenger P* series), *home computers* (as the *Commodore Vic-20* or the *Acorn BBC Micro*) or *videogame consoles* (as the *Atari 2600* or the *Nintendo Entertainment System*)... the "genealogic tree" of the 6502 extracted from the KB helps to visualize a technology in terms of market products. There are a lot of interesting possibilities to count, graph, map and visualize the KB content and they are useful for a public that is wider than scholars.

Museums have a mission of culture preservation and diffusion: their catalogues, though originally addressed to researchers, are the core information used to design and set-up exhibitions. Publishing catalogues on the web[9] is part of an effort to engage a wider audience and to build public awareness and knowledge about HIST cultural heritage.

---

[7] http://www.tnmoc.org/ (last accessed November 13, 2016).

[8] http://hmr.di.unipi.it/CHKB_en.html (last accessed November 13, 2016).

[9] see for instance the Computer History Museum http://www.computerhistory.org/collections/search/ (last accessed November 13, 2016)

With respect to traditional “flat” catalogues, presenting views in the large may be an appealing way to capture the attention of the public and to stimulate its curiosity. Distant reading helps the historians to comprehend scientific and technological history as well as its cultural and social consequences and repercussions. In terms of visualization – such as the technological genealogic trees – distant reading also helps to convey HIST to the public, both as bare facts and as interpretations of the historians – as they can use the views to support their findings.

2.3. Moving in the large

Catalogues of collections are a primary source of information about HIST, but not the only one. There are many other sources: books, media, web pages. The web is a wide and very general container: many books and media are online, at least partially via Google Books or YouTube; there are pages belonging to institutions (some museums have their catalogues online); there are online newspapers and magazines with pages dedicated to technology and HIST is increasingly a topic of interest. There also are pages and blogs written by enthusiasts which, often, offer very valuable information (like deep technical knowledge) mixed with biased opinions (overstated appraisal of a brand or a particular model). Moreover, lot of knowledge is still in the memories of people who were part of the (recent) history of computer industry and is increasingly confided in posts on social networks – again on the web.

Distant reading is usually related to large amounts of data. To extend in the large the contents of our KB it is possible to enrich the information related to collections with additional facts about HIST by extracting them from the web. Browsing the web in search of information worth being inserted in the KB has two other advantages: helps the HIST research and fosters a richer presentation of the history behind the pieces preserved in the museums.

2.4. Additional sources – not always reliable

Feeding the KB browsing the web raises the problem of accuracy and reliability of the information. Which is not a flaw of bloggers and enthusiasts only. Sources like newspapers and magazines should be trustworthy. Yet, the need for stunning news results in highly inaccurate information.

Recently, *La Repubblica*, one of the most important Italian newspapers, titled “The first electronic music track revived, created by Alan Turing” (Rivive la prima traccia musicale elettronica creata da Alan Turing, in the original Italian title[10]). Unfortunately, it was not the first music played by a digital computer and it was not created by Alan Turing. The true story[11] is that the first *recording* of a music played by a digital computer, which was already known, was digitally remastered. Turing was involved in the project at the University of Manchester that built the computer, but the author of the music program was Christopher Strachey. Probably, La Repubblica, as many others, copied the scoop from another (unreliable) source[12], in any case many people were induced to believe that Turing was also the first digital musician of the history.

---

[10] http://www.repubblica.it/tecnologia/2016/09/26/news/turing_musica_computer-148561122/ (last accessed November 13, 2016)..

[11] see for instance http://blogs.bl.uk/sound-and-vision/2016/09/restoring-the-first-recording-of-computer-music (last accessed November 13, 2016).

[12] see for instance https://www.theguardian.com/science/2016/sep/26/first-recording-computer-generated-music-created-alan-turing-restored-enigma-code which correctly cites “recording”, but uses “Enigma” as a (completely uncorrelated) catchword in the URL (last accessed November 13, 2016).

The above example depicts a serious problem we have to cope with, especially if we want to feed the KB with information automatically extracted from the web: it is difficult to discern good information. Sometimes trustworthy sources, like renowned national newspapers, publish false news; moreover, replication of the same information on the web, which usually follows the news, does not work as an indicator of reliability.

2.5. Estimating the KB size

Following a *Fermi problem* approach, we can attempt an upper-bound estimation of the size of the knowledge about HIST. The main goal is to support the feasibility of our solution by giving evidence of a manageable size. We can use the number of *triples* as an unit of measure (and just as that), assuming that triple-store is both a well known paradigm and a quite mature technology with some reference benchmarks (see for instance Oracle 2016). We also consider the knowledge intrinsic in the HIST collections: they are our first interest and a huge concentration of information.

The size of the KB can be estimated by the following formula:

$$\textit{catalogue record fields} \times \textit{collection size} \times \textit{collections for country} \times \textit{developed countries}$$

As the number of the catalogue record fields we can take as an example the *ICCD PST* record proposed by the Italian *Central Institute of Cataloguing and Documentation* for the technological and scientific heritage which count (in the last version) 327 fields (of course not all appliable to every object)[13]. To each field can roughly correspond a triple.

As the collection size we can use the 115233 records of the online catalogue of the already cited Computer History Museum – probably the largest one.

As the number of collections in a country we can take the recent survey[14] made by the Italian Association for Automatic Computing which registered 58 collections in Italy, including museums, retrocomputing associations and privately owned ones.

As the number of developed countries, among many candidates, we can use the number of members of the *Club de Paris*[15].

Summing up:

$$327 \times 115\,233 \times 58 \times 22 = 48\,081\,199\,716 \approx 5 \times 10^{10}$$

Which is a quite large number of triples, yet twenty times smaller than the trillion limit of current technologies (see again, as an example, Oracle 2016). Furthermore, our number actully estimates the number of record fields in the case of flat independent catalogues: using a shared KB eliminates replication of information.

3. KB and KB-driven Information Extraction

As we said before, the first step for populating the KB is bringing together all the museums catalogues. This can be done via some importing procedure from collection catalogues whose results are validated through the mentioned peer review process; supplementary information can be added by editors and reviewers from personal knowledge and research.

---

[13] http://iccd.beniculturali.it/getFile.php?id=258 (last accessed January 8, 2017)

[14] http://www.aicanet.it/dettaglio_evento/598532 (last accessed January 8, 2017)

[15] http://www.clubdeparis.org/ (last accessed January 8, 2017)

Adding to the KB the great amount of information contained in the web can be done via *ontology based information extraction* (OBIE) methods. An OBIE system processes unstructured or semistructured natural language text through an ontology driven mechanism and outputs information structured with respect to the same ontology (Wimalasuriya et al, 2010). The ontology which structures our KB can be used to drive the *information extraction* (IE) process.

However, OBIE systems are not able to avoid all the pitfalls of bad information often found in the web – the above-mentioned musician Turing case can be an example that can deceive an OBIE system, as the fact “first computer music created by Alan Turing” is structurally correct.

To improve the reliability of OBIE, we propose KBDIE. We call a KBDIE system one where the KB plays a double role: the ontology on which the KB is structured is used as in a OBIE system, then the contents of the KB (which we assume not empty and validated) are used to check IE output. In practice, the KB contents are used as a reference in a sort of machine learning process.

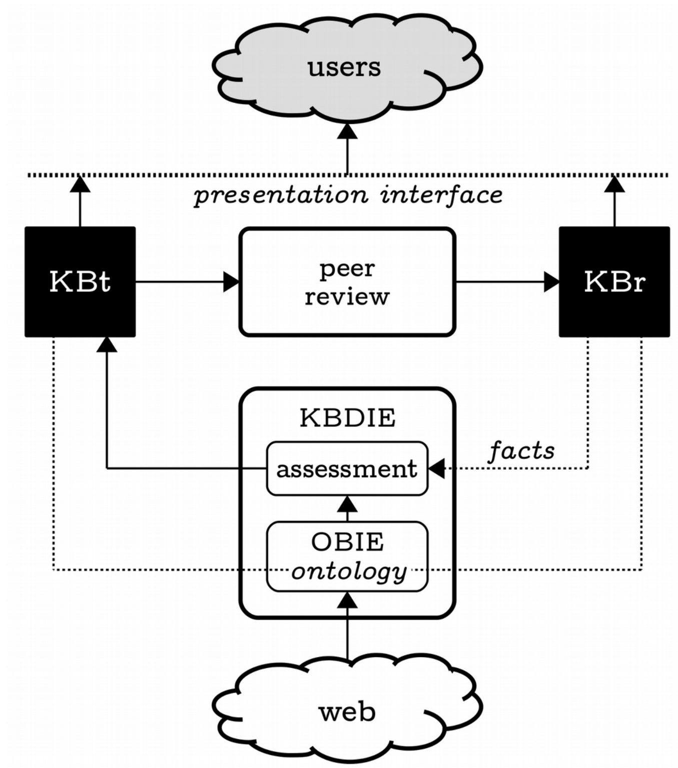

Fig. 1. The KB driven information extraction process

As shown in fig. 1, in our solution the KB is made of two KBs sharing the same ontology. A *reliable KB* (KBr) contains the validated facts that passed the peer review process. A *temporary KB* (KBt) contains only the automatically extracted facts.

The facts in KBr are used to check the IE output and to assess it in terms of *novelty* (that is IE found something that is not in KBr) and *reliability* (what IE found does not negate something which is in KBr). The assessment marks the facts in KBt with scores which help the reader in the interpretation of the KB content. In general, reliability may include other aspects like *temporal validity* and dependance on *personal opinions* – which are lecit. Being an historical knowledge base, time should not be of too much concern. Fans of a particular machine or brand – say for instance the Commodore 64 –

may write on the web biased information. This is harder to spot unless there are evident contradictions with KBr facts – this is one of the reasons for having reviewers in the process.

Back to our musician Turing example, if the KBr contains the fact "first computer music created by Geoof Hill", which refers to the very first music created and played (but not recorded) on the Australian CSIRAC (P. Doornbush, 2005) and/or the fact "first computer music recorded by Christopher Strachey", which refers to the event in Manchester, then the assessment phase of KBDIE can detect a mismatch with the found fact "first computer music created by Alan Turing" and put it in KBt with a "suspicious" score.

KBt facts, even if assessed, cannot go directly in the KBr, as the latter must contain only reliable knowledge. As part of the peer revision process, reviewers examine the KBt facts and move in the KBr those which are considered reliable (after some editing if needed). In other words, automatically extracted information is treated like contents submitted for insertion in KBr as results of the research of keepers, curators and historians in general. To end our example, the fact "first computer music created by Alan Turing" will never go in KBr – unless we consider the idea of a "pit of shame" where bad examples of HIST popularization can be relegated.

Each fact in KBt has the URL of the web source it comes from. The URL can be used in the calculation of the reliability score (depending on previous finding from the same source) and can be used by the reviewers for an additional check. If the fact is deemed reliable and is moved in KBr, the URL is maintained both as a reference and as an indicator of the source trustworthiness.

To the users, KBt and KBr are seen as an unique KB. All the facts are accessible, the difference is in the score, those actually belonging to KBr are marked as "reliable" and, besides the URL of the source (possibly more than one), refer also the names of the reviewers.

Reviewers, as well as editors that contribute by submitting facts directly in KBt, do a very valuable job. Reviewers, being at the end of the process, will be provided with a big amount of information to validate. The automatic assessment is a valuable filter, but cannot guarantee the same level of confidence that an expert can offer. Yet, the work of reviewers should not be very taxing. We expect that they are keepers, curators and historians which use the KB for their usual work and research. They can use KBDIE to order custom searches as well as to submit to the KB the results of their findings. Furthermore, sharing and being in contact with colleagues is a benefit for everyone.

4. Related works

In many fields of research, history shows a shift from "little" to "big science". Our proposal stems from the idea of moving in the large the management of the knowledge about a particular sector of cultural heritage: the collections about HIST. The benefits regard the ability to look at that knowledge from different perspectives, as well as the enabling of cooperation and sharing processes among scholars – and collaboration plays a big role in research (N. Vermeulen et al., 2013).

In (G.A. Cignoni and G.A. Cossu, 2016) we already presented the general idea[16], discussing the differences with respect to traditional catalogues, like for instance the standards proposed by the Italian

[16] The idea has been also presented and discussed at the AIUCD 2015 Conference (http://www.aiucd.it/digital-humanities-e-beni-culturali-quale-relazione-quarto-convegno-annuale-del-laiucd/, last accessed January 8, 2017) and in a workshop organized by AICA and the University of Verona (http://www.aicanet.it/dettaglio_evento/598532, last accessed January 8, 2017).

Central Institute of Cataloguing and Documentation. In the same work we also discussed the differences with respect to very general approaches of union of many catalogues such as, for instance, the *Europeana* project[17] – which adds standard metadata that may help the information extraction and assessment tools of our proposal.

In the proposed solution, like in all OBIE systems, the ontology which structures the KB is designed by experts, defining a priori the entities of the domain and their relations (D. Wimalasuriya, 2010). Ontologies are also used to give structure to knowledge extracted from specific sources, like in the case of DBpedia (Lehmann et al., 2015), where the single source is also assumed reliable. In the KBDIE method in addition to using an ontology to extract information (from the whole web), we propose to use the facts already in KBr as a reference for novelty and reliability evaluation of the new retrieved information.

Machine learning methods are considered very effective to support IE: induction from data produces better results with less effort with respect to definition of formal rules for a logical reasoning system (V. Tresp et al., 2008). In facts, logical reasoning has not proven effective when applied to the web scale nor suitable to manage uncertain information, which is abundant on the web (A. Rettinger et al., 2012). In our case, machine learning is achieved through constantly improving the contents of KBr and by maintaining information about the trustworthiness of sources.

Several approaches focus on learning ontologies, that is using machine learning to extend and improve the ontologies, which is promising on some fields of application (J. Völker et al., 2008). However, in the KBDIE method we are not interested in obtaining a better ontology, as it is assumed that, given the HIST domain, it is possible for experts to design the ontology in advance and we expect it to be stable in time; moreover, being the KB structured on the ontology, changes in the ontology may imply a reorganization of KB contents, which in our case should be supervised.

In (H. Xie et al., 2016) is proposed a statistical evaluation algorithm, which uses a corpus dataset as a reference for evaluating KB triples. Being our triples like <Apple][; released ; 1977> we cannot rely on average values desumed by many repetitions of the same information with slightly different values.

A different approach is in (M. Wick et al., 2013) where sets of triples are scored depending on their inner coherence. This is an interesting direction for building the score, yet in our case supervision is needed: musician Turing would be erroneously considered valid because the news was repeated by many sources. Once that enough facts are stored in KBr, it is likely that using the knowledge in KBr the inaccurate news would be spotted as suspicious.

To improve scoring of found facts, we can use the category of the source: information extracted from a digital copy of a technical manual is more reliable than the one extracted from an advertising brochure. In (F. Sebastiani, 2002) some methods are described for categorising natural text using unsupervised machine learning. As far as it is limited to sources categorisation, this approach looks promising for our case too.

Being mainly interested in assuring the reliability of KBr contents, we generally prefere a supervised approach to accept new facts in KBr. Apart of the role of the KB as reference for HIST, KBr facts are used to assess the new facts in the KBDIE process. Inaccuracies in KBr will affect the assessment and will cause a general decrease of reliability of the whole system. As noted by (M. Wick et al., 2013),

[17] Europeana Collections, http://www.europeana.eu, (last accessed January 8, 2017)

IE systems, however advanced, can offer a confidence level that is not comparable to the confidence of a human expert.

Acknowledgments

We wish to thank the many involved in previous fruitful discussions. Those involved in the development of the prototype, in particular Giuseppe Lettieri from the University of Pisa, and those currently involved in the proposal of an Horizon 2020 EU project on these themes: Giovanni A. Cossu, Norma Zanetti and Sauro Salvadori from Hyperborea srl, Franco Niccolucci from the University of Florence.

References

Boyle, Danny. 2015. *Steve Jobs* (movie).

Cignoni, Giovanni A., and Giovanni A. Cossu. 2016. *The Global Virtual Museum of Information Science & Technology*. Proceedings of International Communities of Invention and Innovation, IFIP WG 9.7 Conference, New York, to be published in Advances in Information and Communication Technology, no. 491, Berlin: Springer.

Doornbusch, Paul. 2005. The Music of CSIRAC: the Australia's First Computer Music. Champaign: Common Ground.

Dou, Dejing, Hao Wang, and Haishan Liu. 2015. *Semantic Data Mining: A Survey of Ontology-based Approaches*. Proceedings of the IEEE 9th International Conference on Semantic Computing (ICSC), USA: IEEE.

Jänicke, Stefan, and Greta Franzini, Muhammad Faisal Cheema and Gerik Scheuermann. 2015. *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges*. Proceedings of the EuroVis 2015 State-of-The-Art Reports.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef , Sören Auer, and Christian Bizer. 2015. *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. Semantic Web Journal, vol. 6 no. 2, IOS Press.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Oracle. 2016. *Oracle Spatial and Graph: Benchmarking a Trillion Edges RDF Graph* (white paper).

Rettinger, Achim, Uta Lösch, Volker Tresp, Claudia D'Amato, and Nicola Fanizzi. 2012. *Mining the Semantic Web. Statistical learning for next generation knowledge bases*. Data Mining and Knowledge Discovery, vol. 24 no. 3, 613-662.

Sebastiani, Fabrizio. 2002. *Machine learning in automated text categorization*. ACM Computing Surveys (CSUR), 1-47. New York: ACM.

Stern, Joshua Micheal. 2013. *Jobs* (movie).

Tresp, Volker, Markus Bundschus, Achim Rettinger, and Yi Huang. 2008. *Towards Machine Learning on the Semantic Web*. In Uncertainty Reasoning for the Semantic Web I, edited by Paulo Cesar G. da Costa, Claudia D'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewcz, Matthias Nickles, Michael Pool, 282-314. Berlin: Springer.

Tyldum, Morten. 2014. *The Imitation Game* (movie).

Vermeulen, Niki, John N. Parker, and Bart Penders. 2013. *Understanding life together: A brief History of collaboration in biology.* Endeavour, vol 37 no.3, 162-171.

Völker, Johanna, Peeter Haase and Pascal Hitzler. 2008. *Learning Expressive Ontologies*. Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 45-69. Amsterdam: IOS Press.

Wick, Michael, Sameer Singh, Ari Kobren and Andrew McCallum. 2013. *Assessing confidence of knowledge base content with an experimental study in entity resolution*. Proceedings of the 2013 workshop on Automated knowledge base construction, 13-18. New York: ACM.

Wimalasuriya, Daya C., and Dejing Dou. 2010. *Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches*. Journal of Information Science vol. 36 no. 3.

Wimalasuriya, Daya C., and Dejing Dou. 2010. *Components for information extraction: ontology-based information extractors and generic platforms*. CIKM ’10 Proceedings of the 19th ACM international conference on Information and knowledge management.

Xie, Haihua, Xiaoqing Lu, Zhi Tang, and Mao Ye. 2016. *A Methodology to Evaluate Triple Confidence and Detect Incorrect Triples in Knowledge Bases*. Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 251-252. New York: ACM.