Automating the Semantic Publishing

Applying a format-independent and language-agnostic approach for the compositional and iterative semantic enhancement of scholarly articles

Silvio Peroni

Digital And Semantic Publishing Laboratory, Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
silvio.peroni@unibo.it

Abstract. The Semantic Publishing concerns the use of Web and Semantic Web technologies and standards for enhancing a scholarly work semantically so as to improve its discoverability, interactivity, openness and (re-)usability for both humans and machines. Recently, people suggest that the semantic enhancement of a scholarly work should be actually done by the authors of that scholarly work and it should be considered as part of the contribution and reviewed properly. However, the main bottleneck for the concrete adoption of this approach is that authors should always spend additional time and effort for actually adding such semantic annotations, and often they do not have that time available. Thus, the most pragmatic way to convince authors in doing this additional job is to have services that enable the automatic annotation of their scholarly papers by parsing the content that they have already written, thus reducing the total time spent by them to few clicks for adding the semantic annotations. In this paper I propose a generic approach called *compositional and iterative semantic enhancement* (CISE) that enables the automatic enhancement of scholarly papers with additional semantic annotations in a way that is independent from the markup used for storing scholarly papers and the natural language used for writing their content. In addition, I report the outcomes of some experiments that suggest that the approach proposed has a quite good margin of being feasibly implemented.

RASH version: https://w3id.org/people/essepuntato/papers/cise-datascience2017.html

RDF statements: http://www.w3.org/2012/pyRdfa/extract?uri=https://w3id.org/people/essepuntato/papers/cise-datascience2017.html

Keywords: Semantic Publishing, CISE, Genuine Semantic Publishing, Structural Patterns, Structural Semantics, Curry-Howard Isomorphism, Principle of Compositionality

1 Print, Digital, and Semantic Publishing

The scholarly communication domain has been involved in several revolutions concerning the way scientific knowledge has been shared in the past 300 years. Guttemberg's introduction of the print (around 1450) together with the creation of formally-defined groups of scholars (e.g. the Royal Society founded in 1660) have permitted research works to be shared according to a well-known medium, i.e. print volumes. Notable examples of this era are Philosophical Transactions published by the Royal Society (first issued in 1776, and still available for submissions) and The Lancet (first issued in 1823, and currently published by Elsevier). Nothing actually changed in such domain until the introduction of the Internet (considering ARPANET as its first implementation around 1960), where communicating research results could happen in a very fast and effective way by means of e-mails. However, the actual first revolution in scholarly communication, since the introduction of the Guttemberg's printer, happened with the advent of the World Wide Web, or WWW, in 1989. The WWW has permitted the explosion of the Digital Publishing, i.e. the use of digital formats for the publication (e.g. PDF) and distribution (e.g. the Web) of scholarly works. In the last twenty years, the availability of new Web technologies and the reduction of the cost for storage have resulted in an incredible growth in terms of the scholarly material available online and in a consequent acceleration of the publishing workflow. However, it is with the advent of one specific set of Web technologies, i.e. Semantic Web technologies [1], that we have started to talk about *Semantic Publishing*.

Semantic Publishing concerns the use of Web and Semantic Web technologies and standards for enhancing a scholarly work semantically (by means of plain RDF statements [2], nano-publications [3], etc.) so as to improve its discoverability, interactivity, openness and (re-)usability for both humans and machines [4] – something which is very close to the recent proposal of the FAIR principles for scholarly data [6]. First examples of Semantic Publishing regarded the use of manual (e.g. [5]) or (semi-)automatic processes (e.g. [7]) for enriching scholarly works with formal semantics **after** their publication, and by people who, not necessarily, have authored such works.

The misalignment between who authored the original work and who added semantic annotations upon it is the focal point discussed by the editors-in-chief of this journal in this introductory issue. Their point is that, after having experimented with the Semantic Publishing, we should start to push and support what they call *Genuine Semantic Publishing*. This *genuineness* basically refers to the fact that the semantic enhancement of a scholarly work should be actually done by the authors of that scholarly work and it should be included in it since the beginning. According to this perspective, the semantic annotations the scholarly work includes should be considered as proper part of the contribution and treated as such (e.g. reviewed properly).

While the Genuine Semantic Publishing is, indeed, the right way to consider the question, it still needs specific incentives for convincing the authors to spend more time on semantically enriching their articles in addition to all the textual content they have already written. In recent experiments colleagues and I have done in the context of the SAVE-SD workshops, described in [8], the clear trend is that, beside a few who actually *believe* in the Semantic Publishing and even if we made available appropriate incentives (i.e. prizes) for people submitting HTML+RDF scholarly papers, generally only a very low number of semantic statements (if none at all) is specified by the authors. Possible reasons for this behaviour could be the lack of appropriate support (e.g. graphical user interfaces) for facilitating the authors in the annotation of scholarly works with semantic data.

However, I do not think that this is even the main issue that prevents a huge mass of authors from enhancing their papers with semantic annotations. I firmly believe that the main bottleneck for a concrete adoption of Genuine Semantic Publishing principles is *author's available time*. While interfaces may simplify the creation of such semantic data, an author should always spend additional time and effort for actually creating them, and often she does not have that time available. Thus, the most pragmatic way to convince authors in doing this additional job is not to have incentives such us prizes, but rather to have services that do it for them in an *automatic fashion* by parsing somehow the content that the authors have already written, thus by reducing the entire time spent by them to few clicks for adding the semantic annotations.

The identification of which dimension to consider for extracting semantic annotations from a scholarly work is subjective. For instance, one can consider more important to have all the citations in a paper described according to their functions (i.e. the reasons why an author has cited other works), while others can consider more functional to have a formal description of the main argumentative claim of the paper with its evidences. In the past years several tools have been developed for addressing the automatic annotation of scholarly texts according to one or more particular dimensions. However, this kind of approaches is typically tied to specific requirements, e.g. the use of a specific *markup* and/or a particular *language* used for organising and writing the content, that prevent it from being used in broader contexts. The idea is that the aforementioned approaches can work correctly only if used with documents stored in a particular format, such as the SPAR Extractor Suite developed for the RASH format [8], and/or if the text to process is written in a particular language such as English, as happens for FRED [9].

Thus, a possible solution for developing *flexible* Semantic Publishing tools that allow the automatic annotation of a scholarly document should at least be independent from:

- the markup format used for storing the document;
- the language used for writing the document content.

In this paper we propose a generic approach called *compositional and iterative semantic enhancement*, a.k.a. *CISE* (pronounced like *size*) that complies with the aforementioned requirements, which has been inspired by a well-known theoretical Computer Science technique called Curry-Howard isom orphism [10] and by the principle of compositionality proper to mathematics, semantics, and philosophy of language. The idea behind CISE is that it is possible to develop and implement a mechanism that enables the automatic enhancement of scholarly papers – independently from the markup used for storing them and the natural language used for writing their content – by means of an iterative process composed by several steps, where each step is responsible for providing additional semantic connotations of the document components included in such scholarly papers by combining the enhancements obtained as outcomes of the previous steps according to precise rules. While a undebatable proof of the validity of CISE has not been provided yet, some experiments have been run in the past and, according to their outcomes, they suggest that the approach proposed in this article can be implemented with a good margin of success.

The rest of the paper is organised as follows. In Section 2 I introduce the foundational theories that have been used to derive the approach for automating the enhancement of scholarly articles. In Section 3 I introduce CISE by describing its main hypotheses and its potential applications. In Secti on 4 I briefly discuss the outcomes of some experiments that implement the first steps of a larger proposal for a CISE-like workflow for enriching scholarly articles stored in XML formats, and I present future directions of research I will investigate. Finally, in Section 5 I conclude the paper.

2 A pathway from syntax to semantics

The Curry-Howard isomorphism [10] is a principle saying that two different formalisms, i.e. the proof system and the model of computation (such as the lambda calculus), are actually the same mathematical tool presented from two different perspectives. The basic idea introduced by this isomorphism is that mathematical proofs can be written as computer programs and, thus, these proofs can be run, and vice versa. This approach has been very relevant for the Computer Science domain, since it allowed researchers to start to investigate new research fields such as type theory, and to create new programming languages for implementing such isomorphism by design, such as Coq.

While the Curry-Howard isomorphism is a quite important milestones in the context of the Theoretical Computer Science area, it has seen several practical applications in different domains. For instance, the use of the isomorphism in the Linguistics domain has been ground-breaking and it has allowed the introduction of what have been known as categorial grammars [11].

One of the pioneers of this kind of grammars, i.e. Richard Montague, in one of his seminal works [12], claimed that there is no difference, at least from a theoretical point of view, between natural languages and logic languages. The consequence of this observation is that it is possible to use precise mathematical tools for defining both the syntax and the semantics of any natural language (e.g. English) in the same way logicians usually do for their artificial languages.

In particular, in his work, Montague used the principle of compositionality, which states that the meaning of a complex object or expression is determined by the meanings of its constituents plus the rules that have been used to combine them. By applying this principle, Montague defined an approach, called Montague grammar, that allows the definition of a precise syntax of a language by means of a set of atomic syntactic categories – such as *noun*, *nominal phrase*, *propositional phrase*, etc. – that can be combined for creating more complex categories by means of specific operators. In addition, this approach allows one to provide a possible semantic representation of each syntactic category by using formulas expressed in lambda calculus. Thus, the idea is that, if we have a language defined by using the Montague grammar, we can write sentences that are syntactically correct and, by means of their mapping, we also can derive their meaning expressed as logic formulas.

Summarising: by applying the Curry-Howard isomorphism in the categorial grammars, it is possible to define the syntax of a natural language in a way that is able to carry the semantic interpretations of the various constituents of that language at the same time. In other words, it is possible to obtain the meaning of a sentence (defined by logic formulas) by analysing only its syntactic composition. It is worth mentioning that this approach is not merely theoretical, but rather it has been implemented by Computational Linguistics researchers, e.g. in Boxer [13], and it has been already applied in real case scenarios in the Semantic Web domain, e.g. in FRED [9].

Of course, the aforementioned applications of the Montegue's approach are strictly related to the natural language text of the content one has to analyse. In fact, all the syntax-to-semantics mappings are strictly dependant on a specific language, mainly English, and thus they can be used only when one is sure to analyse documents written in such language. However, one of the goal of the approach I am envisioning in this article has a specific premise: that the language in which a scholarly document (e.g. a journal article) is written should not be a constraint for the application of automatic mechanisms for deriving semantics from texts.

The approach proposed by Curry and Howard, and then its application in the Linguistics domain introduced by Montague, can be seen from a more general point of view. We are not obliged to take in consideration only natural languages, but rather we can use two independent formalisms (languages, in the broad term) and to define compositional rules that can relate one to each other. Given this premise, the intuition is that we can, in principle, apply iterative transformations starting from the pure syntactical organisation of the various components of a scholarly document (i.e. the elements expressed in a particular markup language, e.g. an XML-based language, and how they contain each other by analysing existing documents) so as to derive their structural semantics (paragraphs, sections, figures, etc.) and, then, their rhetorics (introductions, methods, material, results, conclusions, etc.), and other semantic representations of document components. The idea is that each of these syntactic/structural/semantic aspects, that we would like to arise starting from a pure syntactic organisation of document markup, can be defined in fact as standalone languages (e.g. by means of ontologies) with their proper compositional rules and functions.

The basic principle that would grant the feasibility of this approach is that, while scholarly documents can be written in different natural languages, the main constituents they contain are in fact shared among the whole scholarly communication domain. Of course the way of presenting a research can vary according to the particular research area we are considering – e.g. Life Sciences papers usually follow a well-defined structure, while Computer Science papers are usually organised in a less conservative way. However, I think that, since we are dealing with research works at large, all these kinds of papers are following generic underlying ways of organising the arguments supporting a research. My hypothesis is that such ways are shared somehow between the various research areas.

3 Compositional and iterative semantic enhancement of scholarly articles

Taking inspiration from the ideas introduced in Section 2, the intuition is that, restricting the possible inputs to the sole scholarly papers available in a reasonable and structured markup language (e.g. XML), it is possible to implement a mechanism that can retrieve additional high-level semantic connotations of scholarly papers components starting from their specific low-level conceptualisations. This approach, I have named as *compositional and iterative semantic enhancement* (CISE) of scholarly articles, is thus based on the following general hypotheses:

- [**hierarchical markup**] the sources of the scholarly article are available according to (or can be easily converted into) a document markup language that is appropriate for conveying the typical hierarchical containment proper to scholarly documents (e.g. `body > section > paragraph`);
- [**language agnosticism**] there is no need of having a prior knowledge about the particular natural language used for writing the scholarly article;
- [**layer interdependency**] a layer describing a particular conceptualisation of the components of scholarly documents is dependent somehow on the conceptualisation of at least another lower- or higher-level layer;
- [**inter-domain reuse**] several of the structural and semantic aspects typically described in scholarly articles are shared across various research domains;
- [**intra-domain reuse**] scholarly documents of a specific domain always share several structural and semantic aspects between them, even if such aspects are not implicitly adopted by other external domains.

Intuitively, following the sketch illustrated in Figure 1, starting from a low-level definition of the structure of an article, e.g. the organisation of the XML elements that have been used to describe its content (layer 1), it is possible to create rules that describe each XML element according to more general compositional patterns depicting its structures that oblige a specific implicit and unmentioned organisation of the article content (layer 2) – e.g. the fact that there an element can behaves like a block or an inline item. Again, starting from the definitions in the first two layers, it would be possible to characterise the semantics of each XML element according to fixed categories defining its structural behaviour, e.g. paragraph, section, table, figure, etc. (layer 3). Along the same lines, starting from the aforementioned layers, it would be possible to derive the rhetorical organisation of a scholarly paper, e.g. identifying the argumentative role of each section introduced in such paper such as introduction, methods, material, experiment, data, results, conclusions, etc. (layer 4). And so on and so forth.

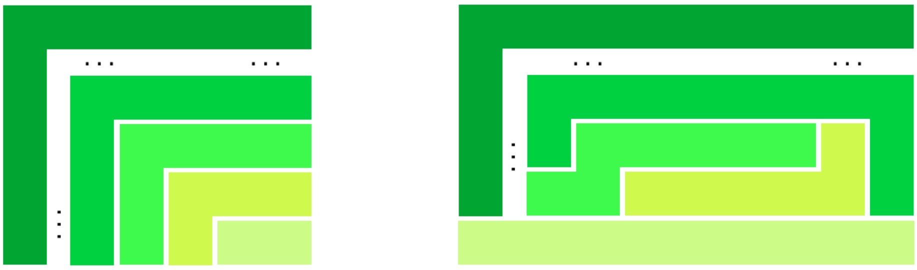

Figure 1. Two graphs depicting possible implementations of the CISE approach. The more the colour of a layer is light, the less the semantic specification is conveyed by that layer. In the graph on the left, the implementation follows a tower-like approach, where a particular layer is totally derived from the information available in the previous one. In the graph on the right, on the other hand, each layer can be derived by means of the information made available by one or more of the preceding layers. It is worth mentioning that these are only two examples of possible implementations of the CISE approach.

Implementations of CISE would allow us to study intra- and inter-domain patterns and similarities that exist between scholarly documents at large. In particular, one of the main aims is to identify which particular structures and semantical connotations are share among different research areas, and how much argumentative information is actually hidden behind quite simple syntactic structures. From such analysis, we could also identifying particular organisation of document components that would allow us to clearly identify paper as belonging to a certain domain, the types of papers (research paper, review, letter, etc.), and other possible ways of discriminating and organising them in clusters sharing similar features. In the next section I introduce first evidences derived by a particular implementation of CISE I have developed with other colleagues in my research group, so as to demonstrate the feasibility of the approach.

4 First evidences and future directions

In the past years I have experimented extensively, with other colleagues in my research group, possible paths for implementing the approach depicted by CISE. Our goal has been to start from a very simplistic description of scholarly articles, mainly related with the pure syntactic organisation of their XML elements, so as to derive more complex semantics by means of an iterative process. In particular, our aims is to retrieve, as outcome of every step of the process, a new layer of semantical conceptualisations of document constituents according to the following scale – from most low-level and pure syntactic description of scholarly articles to their argumentative organisation, where the particular language of each layer is implemented by one specific OWL ontology:

1. syntactic containment (ontology: EARMARK [14]) – describing the dominance and containment relations [19] that exist between the various elements and attributes included in the XML sources of scholarly articles (e.g. the fact that a certain element X is contained in another element Y);
2. syntactic structures (ontology: Pattern Ontology [15]) – starting from the previous layer, inferring the particular structural pattern to which each XML element is compliant with (e.g. the fact that all the elements X behave as inline elements, while the elements Y behave as blocks);
3. structural semantics (ontology: DoCO [16]) – using the particular pattern-element specification provided by the previous layer, extrapolating general rules for the pattern-based composition of article structures (e.g. sections, paragraphs, tables, figures);
4. rhetorical components (ontology: DEO [16]) – by means of the organisation of the structural components obtained in the previous layer, inferring their rhetorical functions, so as to clearly characterise sections with a specific behaviour (e.g. introduction, methods, material, data, results, conclusions);
5. citation functions (ontology: CiTO [17]) – using the outcomes of the previous two layers, assigning the appropriate citation function to each occurrence of a citation in a scholarly articles (e.g. by specifying the function uses method in for all the citations included in the method section);
6. argumentative organisation (ontology: AMO) – analysing the various semantic characterisation of the previous layers, creating relations among the various components of scholarly articles by using specific argumentative models such as Toulmin's [20];
7. article categorisation (ontology: FaBiO [17]) – by looking to the ways document components are organised structurally and argumentatively, annotating each scholarly paper with the appropriate type (e.g. research paper, journal article, review, conference paper, demonstration paper, poster, opinion, report);
8. discipline clustering (ontology: DBpedia Ontology [18]) – understanding to which scholarly discipline each paper belong to by looking at the various characterisations that have been associated to each paper in the previous layers.

The identification of all the information related to the aforementioned layers, while their description could seem easily understandable, is a quite complex work of analysis and derivation and, thus, it is far from being simple and finished. However, in the past six years, colleagues and I have experimented implementations of the aforementioned approach, and we have defined successful pathways for getting from layer 1 to layer 3. From my perspective, these works (described in the following subsections) are clear evidences that the whole implementation of CISE is indeed possible and its principles are valid, at least to certain extend, and that the automatic enhancement of scholarly articles by means of Semantic Publishing technologies can be reached without necessarily using tools that rely on natural language prerequisites or specific markup constructs.

4.1 From containment to structural patterns

The idea of understanding in which way the structure of scholarly documents can be segmented into smaller components, so as to manipulate them independently for different purposes, is a topic colleagues and I have studied extensively in the past. The main outcome of our research on the topic, described in [15], is the proposal of a theory of structural patterns for digital documents at large that are sufficient to express what most users need for writing scholarly papers in terms of document constituents and components.

The two main aspects related to such patterns are:

- orthogonality – each pattern has a specific goal and fits a specific context. It makes it possible to associate a single pattern to each of the most common situations in document design. Conversely, for every situation a designer encounters in the creation of a new markup language, the corresponding pattern is immediately selectable and applicable;
- assemblability – each pattern can be used only in some contexts within other patterns. This strictness provides expressiveness and non-ambiguity in the patterns. By limiting the possible choices, patterns prevent the creation of uncontrolled and misleading content structures.

The basic idea behind this theory is that each element of a markup language should comply with one and only one structural pattern, depending on the fact that the element:

- can or cannot contain text (+t in the first case, -t otherwise);
- can or cannot contain other elements (+s in the first case, -s otherwise);
- is contained by another element that can or cannot contain text (+T in the first case, -T otherwise).

By combining all these possible values – i.e. ±t, ±s, and ±T – we basically obtain eight core structural patterns (described in the Pattern Ontology), namely:

- inline [+t+s+T];
- block [+t+s-T];

- popup [-t+s+T];
- container [-t+s-T];
- atom [+t-s+T];
- field [+t-s-T];
- milestone [-t-s+T];
- meta [-t-s-T].

In [15] colleagues and I introduce a particular algorithm that assigns patterns to the elements of the XML sources (appropriately represented in EARMARK [14]) of scholarly documents without relying on any background information about the vocabulary, its intended meaning, its schema, and the natural language in which they have been written. In particular, this algorithm is the implementation that addresses the passage between layer 1 and layer 2 of the CISE approach introduced above in this section.

The overall process described by the algorithm works as follows. First, it associates an initial structural pattern to each element included in each document available, trying to achieve in-document local coherency of the various pattern assignments according to specific composition rules defined in accordance with the aforementioned pattern theory. Then, the algorithm tries to achieve global coherency between all the XML sources written according to the same markup model. If this is not possible it stops prompting the user to identify possible partitions of the dataset.

The general goal of the process is to understand to what extent patterns are used in several sets of documents, in particular when the markup language used for storing the document content (e.g. TEI [22] and DocBook [21]) is not inherently pattern-based – i.e. it is possible to write a document by means of that markup language in a way that deny a coherent pattern assignment for all the elements described by such language. As consequence of the experimentation of this algorithm with several different corpora of XML sources, we have reached the following conclusions:

- in a community (e.g., a conference or a journal) that specifies the use of a very extensive, permissive, and non-pattern-based markup language, the large majority of authors use a pattern-based subset of such language for writing their scholarly documents;
- only a small number of pattern-based documents coming from different communities (written by different authors in different formats) is needed for automatically generate qualitatively good and generic syntactical visualisation for all documents included in all the communities in consideration.

It is worth mentioning that, once the algorithm has identified that all the elements X in pattern-based documents of a certain community comply with a particular pattern, then it is possible to implicitly associate the same pattern also to all the elements X contained in non-pattern-based documents of the same community. As a consequence, these assignments can also provide a guide to authors (or even to automatic tools) for adjusting the current organisation of non-pattern-based documents so as to convert them in proper pattern-based ones.

4.2 From structural patterns to structural semantics

The systematic use of the patterns introduced in Section 4.1 allows authors to create unambiguous, manageable and well-structured documents. In addition, thanks to the regularity they provide, it is possible to perform easily complex operations on pattern-based documents even when knowing very little about their vocabulary. Thus, the intuition is that it is possible to implement more reliable and efficient tools and algorithms that can make hypotheses regarding the meanings of document fragments, that can identify singularities and, that can study global properties of sets of documents.

Starting from these premises, the algorithm described in a work I was involved in [23] aims at proposing a new algorithm for implementing the transition between layer 2 and layer 3 as sketched above. The experiment introduced in this work involved all the documents of a specific community, i.e. the XML sources of all the papers published in the Proceedings of Balisage (http://balisage.net) encoded in DocBook format [21], that have been already used in the algorithm introduced in Section 4.1. Thus, starting from the outcomes of the aforementioned algorithm, colleagues and I have defined additional rules so as to retrieve the actual structural semantics of XML elements (paragraph, section, figure, table, figure box, table box, bibliography, bibliographic reference, footnote, etc.) by simply looking at the structural patterns that have been assigned to them. The Document Components Ontolo gy (DoCO) [16] has been used for annotating the elements of all the (pattern-based and non-pattern-based) documents with the appropriate semantic structure.

While the experiment run has been done by using documents written a specific markup language (DocBook) and, thus, the algorithm and the results are still preliminary, they are a clear exemplification of how to implement the second enhancement step of the CISE approach as depicted in the beginning of Section 4. In addition, as consequence of this experimentation, we have reached the following conclusion:

- only a small number of pattern-based documents of a community (written by different authors) is needed for extracting the structural semantics of the main components of all the documents included in a particular community.

4.3 What to do next

The pathway for implementing the whole layers of the CISE approach introduced at the beginning of Section 4 is far from being completed, and additional studies and tests are needed for having more robust outcomes. However, I think that the first experiments performed and described in Sectio n 4.1 and Section 4.2 are acceptable pointers for claiming that the automatic semantic enhancement of scholarly articles by means of Semantic Publishing technologies is possible to some extent, even in presence of specific constraints such as the independency from the natural language used for writing such articles and the markup language used for storing their contents.

Some of the other aspects related to the particular implementation of the CISE approach that I intend to study in the future aim at collecting enough evidences for demonstrating the following hypotheses:

- structural semantics of scholarly document components is a sufficient gateway for identifying rhetorical behaviours of the various document parts, such as section rhetorical functions;
- the citation functions, i.e. the author's reasons for citing a given paper [24], can be identified by looking at the rhetorical function of the textual block (paragraph, section) that contains them;
- the main argumentative organisation of a paper can be derived by analysing the rhetorics associated to sections and other components;
- the way scholarly papers argument and organise their content is somehow shared among several disciplines, regardless the actual interrelation that can or cannot exist between such disciplines.

5 Conclusions

In this paper I have introduced the *compositional and iterative semantic enhancement* (CISE) approach for enriching scholarly papers with meaningful semantic annotations by means of Semantic Publishing technologies. Taking inspiration from existing theories in the Theoretical Computer Science domain, CISE depicts a strategy for developing and implementing a mechanism that enables the automatic enhancement of scholarly papers by means of an iterative process composed by several steps, which builds on the idea of providing additional semantic connotations of document components in a scholarly paper by combining the enhancements already obtained as outcomes of previous steps.

I have also discussed some of the outcomes of past experimentations that implement the first steps of a CISE-like workflow that colleagues and I have introduced for enabling the automatic annotation of document components in scholarly papers according to particular structural patterns (inline, block, etc.) and, then, structural semantics (paragraph, section, figure, etc.). Even if these experiments are not an undebatable proof of the validity of CISE, they can be considered first sketches, which suggest that CISE has a quite good margin of being feasibly implemented.

References

1. Tim Berners-Lee, James Hendler, Ora Lassila (2001). The Semantic Web. Scientific American, 285 (5): 34-43. DOI: https://doi.org/10.1038/scientificamerican0501-34
2. Richard Cyganiak, David Wood, Markus Lanthaler (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. https://www.w3.org/TR/rdf11-concepts/
3. Paul Groth, Andrew Gibson, Jan Velterop (2010). The anatomy of a nanopublication. Information Services and Use, 30 (1-2): 51-56. DOI: https://doi.org/10.3233/ISU-2010-0613
4. David Shotton (2009). Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing, 22 (2): 85-94. DOI: https://doi.org/10.1087/2009202
5. David Shotton, Katie Portwin, Graham Klyne, Alistair Miles (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Computational Biology, 5 (4). DOI: https://doi.org/10.1371/journal.pcbi.1000361
6. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, Barend Mons (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. DOI: https://doi.org/10.1038/sdata.2016.18
7. Andrea Bagnacani, Paolo Ciancarini, Angelo Di Iorio, Andrea Giovanni Nuzzolese, Silvio Peroni, Fabio Vitali (2014). The Semantic Lancet Project: A Linked Open Dataset for Scholarly Publishing. In EKAW (Satellite Events) 2014: 101-105. DOI: https://doi.org/10.1007/978-3-319-17966-7_10
8. Silvio Peroni, Francesco Osborne, Angelo Di Iorio, Andrea Giovanni Nuzzolese, Francesco Poggi, Fabio Vitali, Enrico Motta (2016). Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles. PeerJ PrePrints 4: e2513. DOI: https://doi.org/10.7287/peerj.preprints.2513
9. Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, Misael Mongiovì. (2017). Semantic Web, 8 (6). DOI: https://doi.org/10.3233/SW-160240

10. William A. Howard (1980). The formulae-as-types notion of construction. In Jonathan P. Seldin, J. Roger Hindley, To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism: 479-490. Boston, MA: Academic Press. ISBN: 978-0-12-349050-6. http://www.dcc.fc.up.pt/~acm/howard.pdf (last visited May 30, 2017)
11. Raffaella Bernardi (2002). The Logical Approach in Linguistics. In Reasoning with Polarity in Categorial Type Logic. Ph. D. Thesis, Utrecht University. http://disi.unitn.it/~bernardi/Papers/thesis-chapter1.pdf (last visited May 30, 2017)
12. Richard Montague (1970). Universal grammar. Theoria, 36 (3): 373-398. DOI: https://doi.org/10.1111/j.1755-2567.1970.tb00434.x
13. Johan Bos (2008). Wide-Coverage Semantic Analysis with Boxer. In Proceedings of the 2008 Conference on Semantics in Text Processing (STEP 2008): 277-286. DOI: https://doi.org/10.3115/1626481.1626503
14. Angelo Di Iorio, Silvio Peroni, Fabio Vitali (2011). A Semantic Web approach to everyday overlapping markup. Journal of the American Society for Information Science and Technologies, 62 (9): 1696-1716: DOI: https://doi.org/10.1002/asi.21591
15. Angelo Di Iorio, Silvio Peroni, Francesco Poggi, Fabio Vitali (2014). Dealing with structural patterns of XML documents. Journal of the Association for Information Science and Technologies, 65 (9): 1884-1900. DOI: https://doi.org/10.1002/asi.23088
16. Alexandru Constantin, Silvio Peroni, Steve Pettifer, David M. Shotton, Fabio Vitali (2016). The Document Components Ontology (DoCO). Semantic Web, 7 (2): 167-181. DOI: https://doi.org/10.3233/SW-150177
17. Silvio Peroni, David Shotton (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Web Semantics, 17: 33-43. DOI: https://doi.org/10.1016/j.websem.2012.08.001
18. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6 (2): 167-195. DOI: https://doi.org/10.3233/SW-140134
19. Michael Sperberg-McQueen, Claus Huitfeldt (2004). GODDAG: A Data Structure for Overlapping Hierarchies. In Proceedings of the 5th International Workshop on Principles of Digital Document Processing (PODDP 2000): 139-160. DOI: https://doi.org/10.1007/978-3-540-39916-2_12
20. Stephen Toulmin (1958). The uses of argument. Cambridge, Cambridge University Press. ISBN: 9780521827485. http://johnnywalters.weebly.com/uploads/1/3/3/5/13358288/toulmin-the-uses-of-argument_1.pdf (last visited May 30, 2017)
21. Norman Walsh (2010). DocBook 5: The Definitive Guide. Sebastopol, O'Really Media. ISBN: 9780596805029. http://tdg.docbook.org/tdg/5.0/docbook.html (last visited May 30, 2017)

22. Text Encoding Initiative Consortium (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ (last visited May 30, 2017)
23. Angelo Di Iorio, Silvio Peroni, Francesco Poggi, Fabio Vitali, David Shotton (2013). Recognising document components in XML-based academic articles. In Proceedings of the 2013 ACM Symposium on Document Engineering (DocEng 2013): 181-184. DOI: https://doi.org/10.1145/2494266.2494319
24. Simone Teufel, Advaith Siddharthan, Dan Tidhar (2006). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006): 103-110. DOI: https://doi.org/10.3115/1610075.1610091