# ClaimSpotter: an Environment to Support Sensemaking with Knowledge Triples

Bertrand Sereno
b.sereno@open.ac.uk

Simon Buckingham Shum
sbs@acm.org

Enrico Motta
e.motta@open.ac.uk

Knowledge Media Institute
The Open University
Milton Keynes MK7 6AA
United Kingdom

## ABSTRACT

Annotating a document with an interpretation of its contents raises a number of challenges that we are hoping to address via the creation of a supporting environment. We present these challenges and motivate an approach based on the notion of suggestions to support document annotation, hoping these suggestions would act as leads to follow for annotators, therefore reducing some of the difficulties inherent to the task. The environment resulting from this approach, ClaimSpotter, is presented. Aspects of its evaluation are also given, using the findings of a study involving a group of participants faced with a document annotation task.

## Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human information processing*; H.3.7 [**Information Systems**]: Digital Libraries; H.5.2 [**Information Systems**]: User Interfaces

## General Terms

Human Factors, Experimentation

## Keywords

Sensemaking, Annotation, Interface, User studies

## 1. INTRODUCTION

Annotating a document with the information it contains has been addressed through a number of projects [10]. In these scenarios, what the document is being annotated with are *facts*, elements of information which do exist in its contents and which do not suffer discussion and debate. Typical examples include the name and the affiliation of a researcher.

We are interested, on the other hand, in the annotation of documents with knowledge resulting from a sensemaking process. Consider a scholarly document. What its salient points are, how it relates to previous works in a community, and how innovative it is might not appear immediately. Carefully reading the paper, identifying the themes of interest, arguing with the position defended, building on the previous papers we, as scholars, have read beforehand, understanding why a particular paper is cited [22], drawing our own connections to other papers which are (or not) cited... all these steps, among others, are involved in an interpretation process, a sensemaking process. Not to mention that this interpretation might change over time as our research interests evolve, and of course that two readers might see different aspects in the same document. This knowledge is clearly very different from a name and an affiliation.

Annotating a document with sensemaking knowledge, enriching it with such knowledge, allowing users to make it their own by adding personal content to it, therefore raises a number of open questions and challenges of its own. In this document, we will present the formalism with which we are capturing and representing this knowledge, and motivate an approach to support its acquisition, based on the identification of *suggestions* and their presentation in an annotation interface. We then turn to an environment illustrating these principles, ClaimSpotter, and present its functionalities. We will also give initial elements of evaluation, by reporting on a user study involving a group of participants being asked to annotate a scholarly document.

## 2. THINKING IN TRIPLES

The Scholarly Ontologies (or ScholOnto) project [4] proposes an approach to represent, in a semi-formal way, knowledge which by essence is open to debate and discussion, as can be found for instance in academic publications. These documents are annotated (or enriched) with the (possibly contradicting) interpretations made by their readers, who in turn become annotators making sense out of their contents. Interpretations are captured with a number of directed triples (or claims) [source concept, relation, destination concept], where the concepts are unconstrained and can be anything from a word to a paragraph, and the relation chosen from a formal ontology of discourse. Concepts are not to be taken in a strict acceptation here: they are maybe closer to the idea of 'tags' or 'objects', rather than 'strictly and carefully crafted elements of knowledge that everybody agrees on'. An optional type can be added to them, to be chosen among 'approach', 'problem' or 'solution' for instance. One can assign as many of these types to a concept as one wants, by creating as many instances of this concept and assigning


*IUI'05*, January 10–13, 2005, San Diego, California, USA.

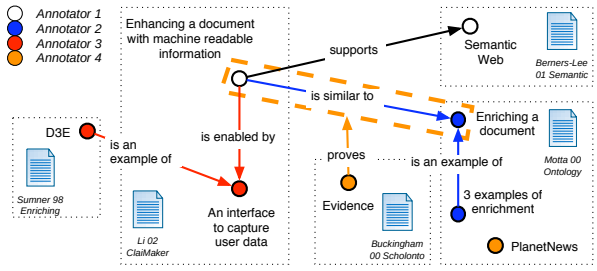


**Figure 1: A network of claims emerges from multiple (and possibly contradicting) interpretations. Concepts are defined for a document and can be reused by different annotators. Claims connect two concepts with one of the relations defined in the ontology. They can also be discussed by putting them at one end of another claim, thus allowing claim-chaining: in this example, *Annotator4* agrees with *Annotator2*'s claim, and adds some *evidence* to strengthen (or *prove*) it.**

the different types to each of them.

The ontology of discourse organizes the way interpretations can be articulated, and contains relations like *addresses*, *proves* or *is consistent with* (see [4] for an account of the different relations available). Figure 1 gives an example of an emerging network of claims (or *claim space*), as it evolves through the multiple (and again, possibly contradicting) annotations of a document's content by different users.

Collecting and compiling such semi-formalized knowledge (for the concepts are free, and only the relations are constrained) enables a number of 'intelligent' uses on top of the network, in addition to the obvious creation of a collective memory. These uses include the identification of the multiple ways to address a particular problem (by following the relations of type *addresses* pointing to a concept, for instance), the identification of groups of people supporting a similar idea, or even the discovery of what we called 'related documents'. Related documents are documents which may not be cited by a given document in its references section, but over which a claim has been drawn by an annotator, connecting a concept defined in each of them. Claims and concepts thus provide another way to discover a relationship between two documents [18]. For instance, in figure 1, the document *sumner98enriching* does not cite *li02claimaker* for obvious chronological reasons, but has become related to it, because *Annotator3* has created a connection between a concept defined in each of them. Annotators can then benefit from this new connection, therefore discovering potentially interesting new documents. Moreover, not only additional documents can be discovered, but the reason, the motivation underlying their connection is also kept.

## 2.1 Difficulties

The semi-formalization also has a cost, of course. We have already mentioned in the introduction the difficulties inherent in the document interpretation process: the content added by annotators to a scholarly document, the content they are enriching a document with, will result from a sensemaking process. Making sense of a document, interpreting a document is difficult. For an interpretation represents a perspective on the document's contents, a vision through a prism reflecting the annotator's own interests, influenced by a number of factors upon which we have no control.

Moreover, formalizing one's interpretation in ScholOnto - in other words, translating it into in a set of concepts and claims - adds this particular problem: the elicitation of what to actually use as nodes in the network (the concepts), how long (or detailed) should they be, and which relation to use to connect them. Translating also means losing some aspects of one's original interpretation, for it will have to be expressed, and articulated, within the constraints of a rigid set of relations. These problems are likely to be faced by newcomers to any application requiring formalization, as noted by Shipman and McCall: *"Users are hesitant about formalization because of a fear of prematurely committing to a specific perspective on their tasks; this may be especially true in a collaborative setting, where people must agree on an appropriate formalism"* [17].

We are not pretending to solve these difficulties, but instead to seek ways to help annotators bridge the gap between their interpretation of a scholarly document expressing the position defended by an author, and the canvas imposed by the ontology of discourse.

# 3. CREATING TRIPLES

We turn now to the approach we have adopted to bridge this gap; in other words to support the underlying sensemaking process and the elicitation of the *bricks* (*id est*, the concepts) to use, and of the ways to connect them (with relations).

The motivation for supporting this process is, as we have just mentioned, its inherent difficulty. One has to get some experience into chunking one's interpretation into concepts, into knowing what chunks make good concepts (according to oneself), and to learn about the different ways available to articulate these chunks. This has to be done in addition to making sense out of the document. The support we are hoping to bring is of course entirely optional. For one because we are not willing to 'force' an interpretation in any way: the different, contrasting, views of two annotators are very welcome. But also because an expert with the formalism could be willing to simply dismiss what we will try to bring her.

Our approach to support the annotation process relies upon the following question:"*what sources of knowledge do we have, and which elements of the document can we find and suggest to annotators ?*" Or to rephrase it, "*how far can we go with text-analysis techniques, bearing in mind that we are dealing with knowledge resulting from a sensemaking process, id est, with knowledge that might not appear in the document itself.*".

The motivation for answering this question is twofold: (1) first of all, by proposing suggestions from the document, we are hoping to facilitate an initial, 'simple', annotation of the document (including for instance, concepts and claims about its contributions, the problem it is tackling, the approach it has used, and so on). Such claims would later be complemented in a 'more complex' annotation, in which annotators would expand their own private claim network by adding new documents (being cited by, or citing, the first document, but not necessarily) they are interested in, and modeling the connections between them; and (2) we are very interested in seeing if these suggestions can make annotators react, and maybe make them *talk* about aspects they had not thought about in the first place. We would like to see if these suggested resources give them additional ideas. To summarise, we are hoping that the suggestions would both get the annotators started, and maybe even spark their interest, by making them react to aspects of the original document they might have overlooked.

## 3.1 Suggestions

Following a strategy in two steps, we consider firstly some characteristics of the document genre we are working with to identify broad areas; we are then combining these areas with the identification of 'information bricks', precisely delimited elements.

### 3.1.1 Relevant areas

We recognize that the nature, the genre of scholarly documents will influence annotators in their concepts and claims formulation task. Studies have shown for instance how people were approaching scholarly documents by breaking them into components [1]. A first aspect of this genre lies in the structure found in scholarly articles. Because of this structure, and because of the annotators' knowledge of this structure (we are assuming here that annotators will be researchers themselves), they will have an idea of what to expect from different components, such as the abstract, the introduction, or a 'related work' section. The author's statements about a particular problem, and the approach she is about to propose to address it, could be found in an *Introduction* section. It would make this particular section worthy of consideration for creating a claim using a relation *addresses* for instance. Similarly, a section '*related work*' would be likely to contain the position of an author with respect to the work of her colleagues. It would make this section a good place to start with for understanding (and potentially modeling, *if* the annotator is interested in these) the connections between the current document and the ones it cites. For these reasons, the ability to navigate within the document and focus on a particular section is desirable.

Although providing access to sections gives a first approximation of the areas where an author would defend her argument, there are also ways to reduce the document to a finer level, by performing further analysis of the contents of these sections. John Swales' account of the rhetorical moves found in the sentences of his corpus of physics research articles' introductions (the Create A Research Space (CARS) model) starts with the hypothesis that the work being described in a research article is expected to be a valid contribution to science. Consequently, authors have to continually defend their position and their contributions, and relate them (through praises or criticisms) to the positions of their colleagues. Argumenting to convince is thus an essential activity of any researcher. His model shows that authors present (and justify the existence of) their work by addressing three different needs: the need to re-establish the importance of the research field; the need to identify a niche where the contribution will be accepted; and the need to occupy and defend this niche [20].

Extending this idea of the role played by a sentence to the entire article, an approach to divide a document into rhetorically-coherent blocks (blocks of text playing a particular role with respect to the functional role of a research article) has been described in [21]. The assignment of a role to a sentence is based on the presence (among other aspects) of linguistic cues such as meta-discourse expressions [12], that are statistically correlated to the rhetorical categories (assigned by a group of experts) of the sentences in a corpus of training documents. The rhetorical roles considered are divided into two broad families: (1) the roles referring to the work being described: these include BACKGROUND statements (attributed to the field in general), the AIM of the paper, its TEXTUAL structure, and other statements which do not fall in these categories (OWN); and (2) the ones indicating a relation to other researchers' work (mostly through citations): these can include a CONTRAST with a previous work, a use of previous work as a BASIS, or OTHER relations which do not fit in these two categories. Sentences tagged with AIM could be proposed to an annotator to make her think about the problem being addressed (resulting maybe in an *addresses* claim). Similarly, sentences tagged with BASIS or CONTRAST could help by indicating to an annotator how the author sees her work in relation with the existing literature. The annotator could then decide to model this relation if she agrees with the author, counter-argue with the author, or ignore it.

Such approaches would help, we believe, the annotator focus on potentially interesting areas. There are of course many other approaches that might be of interest to at least one annotator. We had to start somewhere, and we do believe that the rhetorical parsing approach is the most promising one.

### 3.1.2 Information bricks

We would also like to enhance the suggestion of these areas with an orthogonal, and maybe more ScholOnto-aware, view of the document, based on the identification, the discovery of information bricks, elements which might be valuable enough to be used as concepts (albeit at the price of slight edits from the annotators). By doing so, by proposing tentative concepts, we are hoping to lessen the formalization problem [17] by simply asking annotators whether they make satisfactory concepts according to them, to their interpretation. However, the unconstrained nature of the concepts means that anything, from a word to a paragraph can be a valid concept, and of course, that annotators are still free to create whatever concept they want. We do not want to remove this freedom.

In the case of scholarly document, suggesting the keywords identified by the authors provides an initial set of tentative concepts. More possibilities can also be found by looking for noun groups, sequences of determiners and adjectives terminated by one or several common nouns. They can provide interesting concepts, as well as the acronyms and proper nouns, if any is found.

Another element to look at are instances of the ScholOnto relations found in the text. The motivation here is that the sentences containing these relations (for instance, "The following sections describe three methods developed *to address* these issues." [14]) could provide valuable information by indicating where an author defends her argument, assuming that these relations will be used in their 'academic' acceptation. They could be used as an indication of the areas where an author is, in a way, stating her ScholOnto claims. Annotators would be free to model them if they agree, or take position with them.

Interpretations are also shared and accessible to any annotator. They could offer a valuable resource when annotating a document: the concepts fellow annotators have retained and the claims they have drawn between them are valuable sources of information. Cited and related documents which exist in the ScholOnto knowledge repository could also be accessed and their concepts and claims presented for consideration.

These paragraphs have given an overview of the suggestions we could consider and propose to an annotator to help her bridge the gap between the position defended by the author and her interpretation as a set of concepts and claims. We are now considering the second step: *how* to present these suggestions.

## 3.2 Presentation

Our approach so far suggests an interface where the document would be displayed and browsable. We would also need a way to access and display these suggestions. We are proposing to do this by modifying (where applicable) the document's view, by hiding or emphasizing relevant elements of its contents. Additional resources would be accessible through external windows. The interface should also include a way to record concepts and claims which, upon submission to the ScholOnto database, become additional sources of knowledge to consult for further annotations.

Figure 2 summarizes this data flow. It shows a number of plugins (in the middle column) finding and extracting suggestions, which can be run over the two sources of knowledge we are considering, the original scholarly document, and the repository of concepts and claims (left). The output of these plugins is used in an annota-

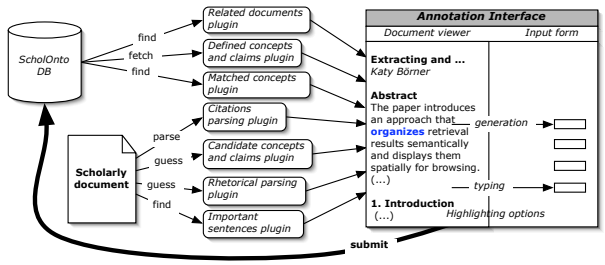


**Figure 2: Architecture of an annotation environment supporting scholarly document sensemaking. Plug-ins extract suggestions from the ScholOnto repository of annotations and from the scholarly document. These can be used to tweak the original document's view, in the hope of providing support for creating concepts and claims. The annotations are stored in the repository upon submission, thus becoming available to the next user.**

tion interface (right). This interface would be schematically composed of two frames: a left frame displaying the document's contents, potentially modified by the output of the plug-ins; and a right frame where annotators would input their concepts and claims. Facilities would be provided to quickly and easily turn some of the suggestions into concepts and claims.

This architecture is also modular, as the plugins are relying on a standardized way to exchange information. This should allow the addition, or the replacement, of suggesting modules, to suit one's needs. With these elements in place, we now turn to the application we developed to illustrate these aspects.

# 4. ANNOTATING WITH TRIPLES

The ClaimSpotter interface is the latest member of a suite of interfaces we have been developing over the last years to annotate documents with concepts and claims, but is the first one considering suggestions extracted from the text and from the repository to support this process. It is a Web interface, principally acting as both a document reader [9], where the appearance of this document can be modified by the activation of different plug-ins, and a form to record concepts and claims. Figure 3 displays a typical work session with ClaimSpotter.

On the left side of the document panel, a 'table of contents' panel permits a quick navigation within the document. On the right side is the form panel where annotators can input their concepts and easily combine them in claims. A toolbar gathers all the different suggestions one can activate to modify the document's view, including: (i) the concepts made by the current annotator or by every annotator; (ii) the instances of the ScholOnto relations (verb expressions) found in the text; (iii) the document's 'important' sentences (where the importance of a sentence is assessed trivially by checking whether each of its words appears in the title, the headers or the abstract); (iv) its rhetorically-consistent zones (using a home-grown and much simplified version of the tool described in [21]); and finally, (v) the sentences matching a particular user-defined query expression. These elements are discovered through our initial implementation of the plug-ins presented in figure 2. We should stress here that the focus was on testing their relevance and their integration in an interface. The modularity of the architecture means that it should not be too difficult to replace them with more robust iterations. It is indeed planned to replace, for instance, our own basic rhetorical parser with a state-of-the-art version as found in [21].

By using and combining these suggestions, an annotator can build her own representation of a document, thus showing or hiding as much as she wants. In figure 3, the user has combined 'her concepts' (the concepts she has defined previously, in any document, matched in the current document's text), instances of the ScholOnto relations, sentences with an importance score higher than 5 and the sentences containing the term 'Trellis'. We believe the ability to manipulate the document the way one wants, and getting as many (to reduce cognitive overload) or as few suggestions as wanted is a crucial aspect of this task.

Concepts can be typed in the form panel directly, dragged and dropped from the document's contents, or reused by clicking on their occurrences in the document (in the case of concepts already existing). In the latter case, their history can be accessed, to reveal who created them and how they have been used over the corpus. An additional 'get more ideas' window provides a number of tentative concepts extracted from the text (document's keywords and noun groups), and facilities to import them with a one-click operation.

Claims can be made by typing them in directly, or by combining concepts (again, in a one-click operation) created earlier and selecting the relation to use to connect them. Clicking on an instance of a ScholOnto relation in a sentence creates a claim by splitting this sentence in a triple centered on the relation. It is also possible to import a claim recorded earlier, by any annotator, if the current annotator wants to state it again, strengthen it or take position with it (an example of this scenario was given in figure 1).

A notepad is also incorporated, allowing annotators to scribble down ideas, to help them in their task. These notes can be parsed to look for instances of ScholOnto relations and selected Wordnet synonyms. If a relation is found, the contents of the note is split and a claim is created, centered around this relation.

The history window can be used to create a summary view of any document (showing its associated concepts and claims, and offering the possibility to reuse them for the current document), concept (to discover *who has created it, which document has it been used in, and by whom*, and *which claims is it involved in ?*), or claim (*which concepts are involved in it* and *has this claim been discussed and possibly challenged in another claim ?*). Cited and related documents, and their associated concepts and claims, are also accessible from this window.

A search facility allows a quick navigation of the underlying repository, to find any concept, claim or document title matching a particular query expression.

The concepts and claims submitted by an annotator are immediately visible in the document view (by highlighting one's existing concepts matched, or everybody's), and in the history window. New documents might also become related because of these newly submitted concepts and claims (we explained how in section 2), and the annotator can access their information immediately.

An export option can be activated to generate a graphical map of the annotation of any combination of document(s) and/or user(s), similar to the one presented in figure 1.

The motivation behind this environment and its underlying architecture was to support annotators through the extraction and presentation of suggestions, providing an environment where they could consult, combine and discuss multiple ressources. We were hoping to make this sensemaking and formalization process a tad easier for them, and, as we mentioned earlier, to spark their interest by making them think about aspects of the documents (highlighted by the suggestions) that they might have overlooked. The user study that we will present now will give us initial elements of answer to assess whether this approach was successful.

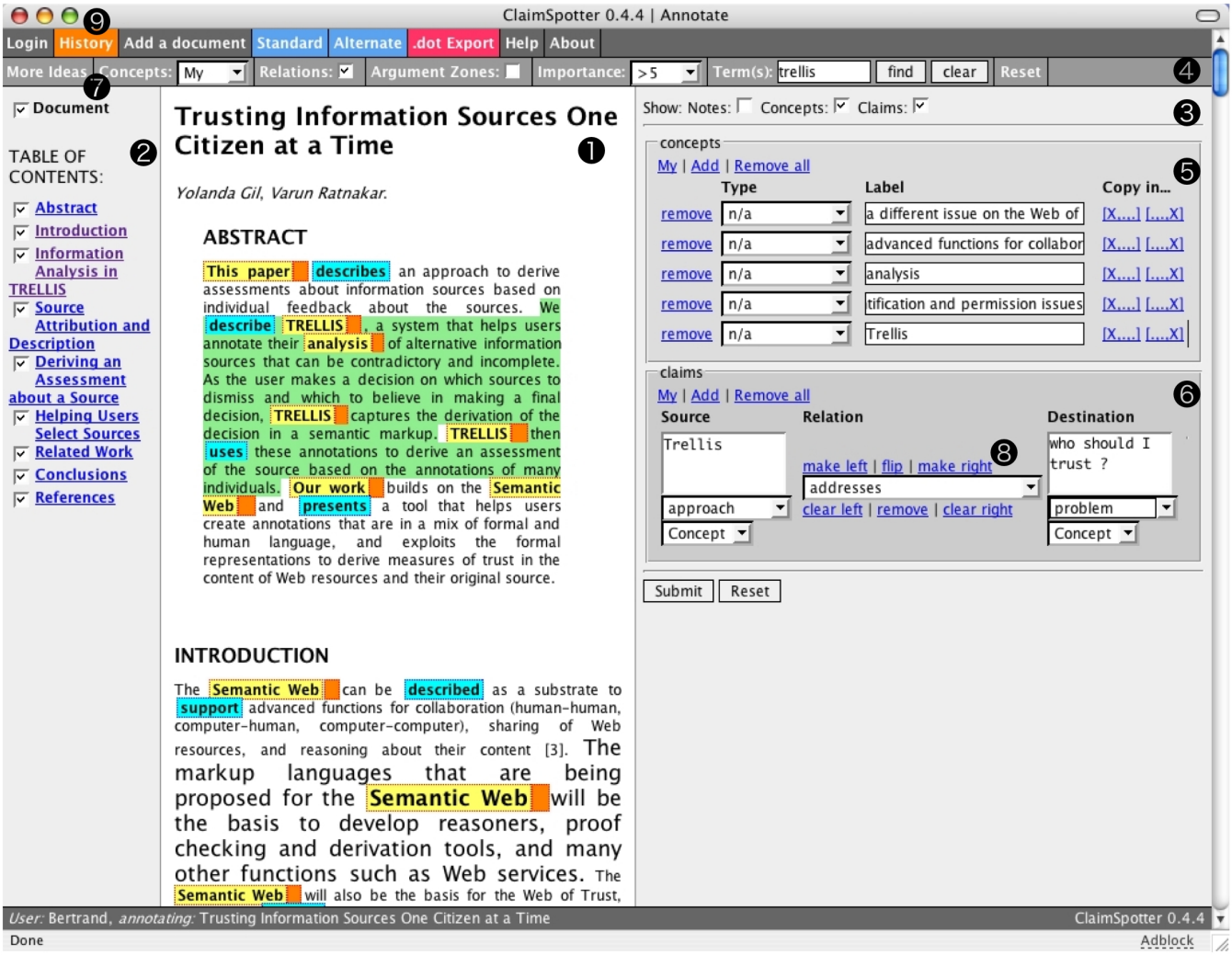


**Figure 3: Using ClaimSpotter: an annotator has combined a few filters ④ on the document panel ① to get suggestions and find aspects to consider and/or react against. The structure of the document is presented ② to permit a quick navigation. Elements found in the text can include existing concepts, the relations defined in the ScholOnto ontology (or verb patterns) or sentences matching a user-defined query. The form panel ③ allows the insertion of notes, concepts ⑤ and claims ⑥. Concepts (imported from the text, from the 'More ideas' window ⑦, or typed) can be combined in triples by choosing the desired relation ⑧. The History window ⑨ gives access to the contents of the repository and provides summary views for each document, claim and concept available.**

## 5. USER STUDY

The user study was driven by the following questions: (1) *how usable the interface is ?*; and (2) *how (and if) the interface and the suggestions influence the strategy of the annotator ?*

We have studied the interactions of 13 users of ClaimSpotter. Their behaviour on screen, and their voices were recorded in movies, one by one. This group was composed of 4 ScholOnto experts (members of the project's team and familiar with the relations available), and 9 beginners. Each of them was asked to annotate a short paper - that either they had written, or that they were familiar with - with concepts and claims. This decision was made because in a real life setting, annotators would of course only annotate the documents they are interested in. Asking them to be familiar with the document was also reducing the cognitive overload they would probably face, having to deal with a new interface, and for the beginners, a new formalism.

Their task was therefore to enrich a document, to add their content to it in the form of concepts and claims connecting these concepts. Each participant was given an initial tour of the interface and of its functionalities. Beginners were also given an additional presentation of ScholOnto before the experiment took place. They were asked to prepare the annotation before coming in for the experiment, by identifying concepts and claims on their own, without the help of the ClaimSpotter environment. An additional questionnaire was sent later, after the experiment, asking them their general opinion about the tool.

Figures are not relevant in this scenario, for the number of concepts and claims one submits is in no way an assessment of the quality of an annotation: meaningful interpretations can be done in a few salient concepts and claims, while a high number of concepts and claims can discuss mundane aspects. However, the very fact that all the participants managed to 'get something done' seemed

to indicate that the environment was performing reasonably well (participants on average inputted 20 concepts and 11 claims each, with no noticeable difference between the experts and the beginners)

## 5.1 Getting support

We were expecting to witness as many behaviours as annotators, and partly this is what we observed. About a quarter of the participants (spread across our experts and beginners) made little use of the suggestions and spent most of the experiment inputting concepts and relations in the right hand part of the screen, while the remainder (once again, evenly spread between experts and beginners) did actually use the suggestions. One expert preferred at a time to deactivate the suggestions '*to keep things simple*', while one beginner said that she liked '*things coming out automatically*'. It seems here that there is no rule governing the users' needs. There will be times when they prefer to have little help, so that they are not influenced by existing suggestions. There will also be times when they will feel a need for suggestions, for instance if they do not know where to start their annotation from. The flexibility of the interface, and its possibilities to suggest as much or as little as desired, seem to cover these needs.

## 5.2 Identifying concepts

The annotation was often done incrementally, by submitting an initial set of concepts and claims, and adding new ones building on or reusing these initial ones. When that happened, participants tended to immediately check the history window and the document to see if their concepts and claims had been recorded.

This is actually showing how the annotation of a document became pushed towards concepts. Annotators did find it very easy to create concepts, and most of the time, these concepts were being created out of copied and pasted text. While there is no way to assess whether concepts created from existing text are better or worse than concepts made 'purely' out of reflection, it is interesting to notice how annotators reacted to the presence of the text. It could be explained by a natural 'laziness'. And by another aspect of the interface: the fact that it colorizes the concepts matched exactly in the document as soon as they are submitted. Annotators could therefore be influenced and tempted to create concepts from an uncolored area because they would like, to quote one annotator, '*to add some color in there*'. We reported elsewhere on how a graphical interface shaped the annotation more towards the articulation of these concepts into claims [6].

Reusing concepts created by fellow annotators can be also seen as a form of reaction towards existing content. It may be that these concepts were representing exactly what the current annotator had wanted to formalize, or that they were '*quite good*' (not exactly matching the original intent, but good enough compared to the cost of creating one from scratch).

## 5.3 Articulating concepts

The most crucial aspect we observed was the difference between how (mostly) experts were starting from the relation they wanted to create ('*I want to say that this concept impairs this one*'; 'impairs' being one of the relations in the ontology), and how beginners, most of the time, were starting from the two concepts they wanted to connect, without knowing if the relation they wanted to use was actually existing in the ontology ('*And this one.. and this one are in a relationship*'). We used the expressions 'mostly' and 'most of the time', for experts did fall in this situation sometimes, and beginners did start from the relation they wanted to use in some occasions too, as they were starting to remember the relations available.

In most cases, independently of this distinction, creating claims required a certain amount of trying and manipulating to get them right. We witnessed how annotators changed their strategy when they could not find a satisfactory relation. Sometimes, it was as easy as flipping the left and right parts to make a relation click in (one of the most well-received features: as one participant said, "*It's a very useful thing because that's something a human would do, not getting my concepts the right way around, because I am thinking and just throwing concepts in.*"). Some other times, they had to actually reformulate the concepts to make a relation fit. When this process was succesful, a claim with a relation matching their original intended meaning was submitted. It also happened that the relation chosen in the end was "*not perfect but kind of all right*". If the claim could not expressed, it was dismissed.

Among the comments we received, one of the most frequent (coming from beginners), was about why a particular relation they wanted to use was missing. A similar comment was also made for the optional concept type one could add to a concept. One type (concept type or relation) was sometimes missing; some other times, a particular type was acceptable but '*not capturing exactly what (I) wanted to say*'. This seems to be a very important aspect. Of course, it is easy to answer to it by saying that annotators have to comply with the ontology of relations, and that this ontology is expressing the view the project members were interested in. Annotating in a formal scheme is indeed translating, and it implies losing some of the original meaning. The very fact that annotators cannot input the claim they wanted to make (no matter if it is simply not possible to do so) can have a consequence on how one approaches the task though ('This tool is not interested in what I want to say.')

## 5.4 Providing more support

These differences in annotators' behaviors are only a subset of the reactions we compiled over the user study. The most striking one would be the difference of approach between experts and beginners, as shown in the 'I start from the relation' vs. 'I start from the concepts' dichotomy. We do not know if much could be done here, as it seems to be related to the knowledge one has of the formalism, and therefore of one's experience with it. A help screen listing the different relations was available, but none of the beginners made use of it. One direction we could follow though, would be to guide annotators more, by providing a walkthrough. Suggesting steps like 'identify the problem addressed in this paper', 'state the contributions and relating them to the problem', 'relate the contribution to the contributions stated in cited documents' and providing support (via suggestions) to answer those points could be useful and help them both in approaching the paper and in realising what the ScholOnto ontology captures. The next section will present our first steps in this direction.

# 6. TOWARDS A PROACTIVE VIEW

We are presenting now our first steps towards a module offering active support for the task of interpreting a document. This work is currently under development and is aiming at giving more support to beginners.

To achieve this, we are proposing to add structure to the way annotators approach the task by asking them a set of questions designed to drive them through the different families of relations available in the ontology, and by proposing answers to these questions, based on extracts from the original document expressing the author's position (if applicable), and on elements drawn from the repository of annotations, expressing the annotators' points of views.

The range of questions being proposed for consideration to the annotators is given in figure 4. Support to answer the first ques-

| | |
|---|---|
| q1 | What is the problem identified in this document ? |
| q2 | How is this problem related to other problem(s) ? |
| q3 | What are the approaches and solutions proposed in this document ? |
| q4 | What are the claims connecting problem(s) and solution(s) ? |
| q5 | Which claims involving the solution(s) are defined in this document ? |

**Figure 4: The current set of questions asked by the (optional) proactive module, aiming at structuring the way one approaches and formalizes her interpretation.**

tion could be gathered by proposing sentences classified as AIM by a rhetorical classifier [21], the concepts defined over this document which have been typed as 'problem', and the destination ends of claims using an *addresses* relation. The module has not been inserted in the main interface yet, but one could imagine a questionnaire standing aside the concepts and claims panel in figure 3.

Another evaluation is also under way for this new module, based on the reaction to these additional suggestions. We have so far only had two participants, but the results already seem interesting. Both agree it would be useful, for the same reason: as a "*walk-through/overview (especially if I am not totally familiar with the document)*", or to "*make you think about a paper and identify its structure*. They expressed concern however about the amount of information that could be presented ("*I wouldn't want to be flooded with everyone else answers.*") and its quality. Filtering options would be needed to ensure annotators get as much, or as little, information as they need.

One annotator also correctly points out that there will be times where she would "*spend time making my own claims.*" Conversely, "*there would (also) be times where extensive reuse of claims is the best approach.*" This is again an interesting comment, for it seems to suggest two different scenarios in the future: the annotation of papers that do really matter to the user (possibly her own papers), in which cases the annotator does have enough confidence (or will) to go down the road on her own. And a different scenario in which the paper to annotate is still important, but maybe not as much, in which case support to make it, should we say, as quickly and as painlessly as possible would be appreciated.

This module seems to be most promising, and its development will be continued over the next months.

## 7. RELATED WORK

Although it has a mostly text-driven interface, ClaimSpotter is ultimately concerned with the creation of argument maps expressing, for instance, the position of multiple annotators over a particular problem [11]. Such maps could be used to represent the perspective taken on a domain, according to the different annotators (and potentially authors) of the documents being connected. Additional work was also carried out on the possible ways to represent and navigate ScholOnto argument maps [5].

ClaimSpotter lies a step behind this, as it is concerned with the initial process by which these maps are created, *id est* the capture of their nodes (concepts) and edges (claims). By supporting the annotation of a document with a semi-formal interpretation of its contents, it shares characteristics with both traditional knowledge-based annotation tools and tools supporting the capture of interpretation (and subsequent discussion) of ressources.

In the former category, tools like CREAM [10] support the annotation of *facts*, elements of knowledge which are not likely to be discussed and debated. It displays the document to be annotated on the right side of the interface, and the list of classes, instances and attributes in a 'control panel' on the left side. An *annotation by markup* mode is provided, where the user selects any piece of relevant information from the page and drags and drops it to create or instantiate the selected element. ClaimSpotter provides a similar drag and drop annotating mechanism. What we are interested in "remembering, thinking, clarifying and sharing" [16], however, results from a sensemaking process and might therefore not appear in the document.

In the latter category, we find tools supporting debate and discourse, as for instance D3E [19]. The document is here displayed on the left side, and a threaded discussion space where reviewers can discuss and argue about the position defended in a document is on the right side. Links from each unit of the document to the appropriate point of the discussion space and vice-versa are provided. No further formalization, apart from an 'agreement' or 'disagreement' tag matching the nature of the comment being submitted is required. The debate and discourse supported in ScholOnto does require formalization, but expects to derive from it a range of intelligent services.

TRELLIS is another system which adds formal structure to the semantic annotation, by linking statements drawn from Web documents, using a set of discourse, logical and temporal connectives [8]. A system to support the incremental formalization of these statements through ontology paraphrasing has been recently added to it [2]. To replace user terms with entities predefined in a domain ontology, an engine looks for matching entries and uses the relations between the ontology entities to propose suitable replacements. This approach is, however, relying on the presence of a domain model articulating the different concepts one can find in it. In our case, we do not have such a model, for the formality is in the relations only and the concepts are free, which limits the support we can bring in assisting the rewriting of user statements (the only similar support we could provide (and actually are providing) would be via the parsing of the free-text notes, where we could look for instance of ScholOnto relations, or of their synonyms, to transform them into claims, if possible).

ClaimSpotter and its underlying formalism combine elements from these two families of applications by providing a knowledge-based annotation tool for concepts and claims, and supporting debate and discourse through the possibility offered to consult and discuss (via claim-chaining, see figure 1) the annotations of fellow users. It supports this process by adding a layer on top of the document, in an attempt to reduce the amount of information and therefore ease the task. This layer is partly comparable to the one described in Magpie [7]. Given a Web page describing a researcher's activities, and an ontology describing the members, activities, projects and research areas of an organization, Magpie can contextualise its content by highlighting the relevant information and providing semantic services on top of this. Our own layer (which does not add semantic services for there is no domain model) adds more elements to a page as it not only shows existing members of an ontology (annotators' concepts and the ontology's relations) but also a number of additional elements (rhetorically-coherent zones and important sentences, for instance). Suggesting information to assist the creation of a knowledge map is also inspired by the work of David Leake *et al* [14]. This approach proposes methods to assist experts and beginners alike in their task of building and extending a knowledge map by adding concepts and connections (or propositions).

Much work has also been done in text analysis techniques adapted to scholarly documents, including work on citation motivation pars-

ing [15] (and [21], which we presented earlier) and citation indexing [3]. We are hoping to include more of these works in a future version of ClaimSpotter.

## 8. CONCLUSION

In this paper, we have stated the difficulties inherent to the task of annotating a scholarly document with a semi-formal representation (as concepts and claims) of its interpretation. We have hypothesized that providing support by suggesting and presenting elements in a supporting application could, at least partly, alleviate some of these difficulties and both get the annotators started and potentially make them say 'better things', by making them react to a set of suggestions. We have presented a number of these suggestions and explained their integration in an open architecture, where additional components could be added if needed.

The subsequent user study has given us some insight as to how a group of both experts and beginners were approaching this task, the difficulties they encountered and the support they found in the suggestions. Interesting differences in their behavior were noted and analysed. We are very interested in trying to provide additional support, especially to beginners, and we have presented the first version of a new module providing a walkthrough guiding the interpretation and suggesting tentative answers.

## 9. ACKNOWLEDGMENTS

This research was partially supported by the Advanced Knowledge Technologies (AKT) project. AKT is an Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University.

## 10. REFERENCES

[1] A. P. Bishop. Digital Libraries and Knowledge Disaggregation: the Use of Journal Article Components. In *Proceedings of the 3$^{rd}$ International Conference on Digital Libraries*. Association for Computing Machinery, 1998.

[2] J. Blythe and Y. Gil. Incremental Formalization of Document Annotations through Ontology-Based Paraphrasing. In *Proceedings of the 13$^{th}$ International World Wide Web Conference (WWW2004), New York City, NY, USA*. Association for Computing Machinery, May 2004.

[3] S. Bradshaw and K. Hammond. Automatically Indexing Documents: Content vs. Reference. In *Proceedings of the 6$^{th}$ International Conference on Intelligent User Interfaces (IUI02)*. Association for Computing Machinery, 2002.

[4] S. Buckingham Shum, E. Motta, and J. Domingue. ScholOnto : an Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, 3:237–248, September 2000.

[5] S. Buckingham Shum, V. Uren, G. Li, J. Domingue, and E. Motta. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter 9, pages 185–204. In Kirschner et al. [13], 2003.

[6] S. Buckingham Shum, V. Uren, G. Li, B. Sereno, and C. Mancini. Computational Modelling of Natural Argumentation in Research Literatures: Representation and Interaction Design Issues. *To appear in the International Journal of Intelligent Systems (IJIS) special issue on Computational Models of Natural Argument*.

[7] J. Domingue, M. Dzbor, and E. Motta. Magpie: Browsing and Navigating on the Semantic Web. In *Proceedings of the 8$^{th}$ International Conference on Intelligent User Interfaces (IUI04), Funchal, Madeira, Portugal*, pages 191–197. Association for Computing Machinery, January 2004.

[8] Y. Gil and V. Ratnakar. Trusting Information Sources One Citizen at a Time. In *Proceedings of the 1$^{st}$ International Semantic Web Conference (ISWC 2002), Sardinia, Italy*, 2002.

[9] J. Graham. The Reader's Helper: a Personalized Document Reading Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI99)*. Association for Computing Machinery, 1999.

[10] S. Handschuh and S. Staab. Authoring and Annotation of Web Pages in CREAM. In *Proceedings of the 11$^{th}$ International World Wide Web Conference (WWW2002)*, 2002.

[11] R. E. Horn. *Infrastructure for Navigating Interdisciplinary Debates: Critical Decisions for Representing Argumentation*, chapter 8, pages 165–184. In Kirschner et al. [13], 2003.

[12] K. Hyland. Persuasion and Context: the Pragmatics of Academic Metadiscourse. *Journal of Pragmatics*, 30:437–455, 1998.

[13] P. A. Kirschner, S. Buckingham Shum, and C. S. Carr, editors. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer Verlag, 2003.

[14] D. B. Leake, A. Maguitman, T. Reichherzer, A. Cañas, M. Carvalho, M. Arguedas, S. Brenes, and T. Eskridge. Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. In *Proceedings of the International Conference On Knowledge Capture (KCAP), Sanibel, FL, USA*. Association for Computing Machinery, October 2003.

[15] H. Nanba and M. Okumura. Towards Multi-paper Summarization using Reference Information. In *Proceedings of the IJCAI'99 Conference*, pages 926–931, 1999.

[16] I. Ovsiannikov, M. Arbib, and T. McNeill. Annotation Technology. *International Journal on Human Computer Studies*, 50(4):329–362, 1999.

[17] F. M. Shipman and R. McCall. Supporting Knowledge Base Evolution with Incremental Formalization. In *Proceedings of the SIGCHI 1994 Conference on Human Factors in Computing Systems conference*, pages 285–291. Association for Computing Machinery, April 1994.

[18] H. Small. Co-citation in the Scientific Literature: a New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

[19] T. Sumner and S. Buckingham Shum. From Documents to Discourse: Shifting Conceptions of Scholarly Publishing. In *Proceedings of the SIGCHI 1998 Conference on Human Factors in Computing Systems*. Association for Computing Machinery, April 1998.

[20] J. M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[21] S. Teufel and M. Moens. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445, December 2002.

[22] M. Weinstock. Citation Indexes. In *Encyclopedia of Library and Information Science*, volume 5, pages 16–40. 1971.