REPRESENTING SCIENTIFIC KNOWLEDGE

Rosina O. Weber and Sidath Gunawardena
Drexel iSchool
3141 Chestnut street Philadelphia, PA 19104 USA

ABSTRACT

A representation for scientific knowledge can enable the computational manipulation of scientific contents. The goal is to facilitate the construction of deliverables in support of education. These learning products can range in a variety of dimensions: in format from textual to computational; in educational content from scientific articles to textbooks; in audience interests, from politicians to the general public; they can be interactive, passive, or proactive. The need to enhance our capacity to build learning products originates from the current multifaceted media context. Educators need to compete for students' interest with sophisticated forms of web content built with the participation of millions. This work proposes the representation and manipulation of units of scientific knowledge, *learning units*, to (semi-)automatically manipulate scientific contents to efficiently construct effective learning products. We illustrate the use of learning units for the automated construction of slide presentations.

KEYWORDS

Knowledge representation, acquisition; learning and expertise; knowledge management; knowledge engineering.

1. INTRODUCTION

Representing scientific knowledge is required to leverage our ability to construct learning products such as texts, videos, and cognitive tools (van Joolingen 1999). Such a representation has to be machine-readable to be amenable to automated methods. The entire approach is multidisciplinary because it uses scientific knowledge from one or more disciplines, knowledge engineering methods for manipulating the representation, and knowledge of multiple disciplines to guide design of the respective product.

These learning products can range in a variety of dimensions. The learning goal can vary from expository to discovery. The student can be from novice to advanced. The target format of the product can be text, video, audio, and computational. When computational, they can be interactive, passive, or proactive.

The need of cognitive tools has been constantly furthered by the exponential rate content is created today. Moreover, technology created a culture where attracting the attention of students has never been this hard (Lanham 2006). Hence, learning tools need to increase in sophistication.

We propose an approach that uses the Learning Units paradigm. Learning Units (LUs) represent scientific knowledge by combining elements of procedural, declarative, and structural knowledge; elements that are studied in knowledge engineering (Weber et al. 2008). The development of LUs is also highly influenced by knowledge management concepts. This representation paradigm of has been successfully used to elicit, share, and summarize knowledge from scientists. LUs guide users to enter only useful knowledge while capturing it in a form that is computationally tractable.

In previous work, we used LUs to automatically generate drafts of reports (Weber et al. 2007). In order to illustrate our proposed approach, here we present how thisparadigm can be used to automatically create drafts of slide presentations.

2. THE APPROACH

This work assumes that a machine readable representation is pre-requisite to the efficient development of multiple education products. Figure 1 demonstrates the context of the approach, which entails three stages,

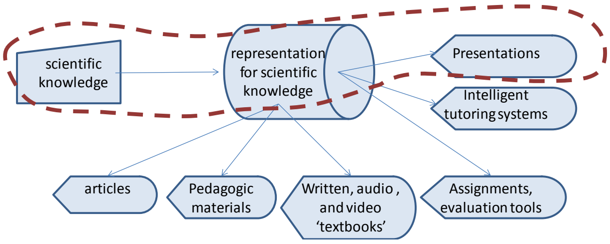

Figure 1. Knowledge - representation - education products

namely, scientific knowledge identified, scientific knowledge converted into a representation paradigm, then represented knowledge used to create education deliverables.

There are multiple sources of scientific knowledge, the primary being scientists. For this reason, LUs are elicited from scientists. In previous work we discussed and demonstrated the process of eliciting scientific knowledge from scientists directly into LUs (Weber et al. 2006, 2008). Scientific knowledge can be also found in publications. The process of extracting scientific knowledge from these textual sources into LUs is an ongoing effort. The machine-readable representation allows its computational manipulation, enabling the automated generation of educational deliverables. This last stage requires knowledge of design of such deliverables. Here this approach becomes highly discipline centric as the design of each deliverable has to use knowledge from the field that utilizes those deliverables.

The range of educational products is virtually unlimited. Some possible dimensions are format, audience, level, and purpose. These products do not have to be limited to an audience of students; they can target audiences from multiple backgrounds: diverse like the general public or specific like politicians and economists. While the design knowledge required in the creation of these products usually comes from humans, it can also be obtained from communications.

The format of target deliverables can be textual or graphic, and include audio and video. This requires knowledge from digital media and graphic design, where as computational tools will require knowledge from systems analysis and computer science. More traditional learning and education products include scientific articles, textbooks, course descriptions, course outcomes, and program curricula. These require knowledge in instructional design.

3. REPRESENTATION PARADIGM: LEARNING UNITS

Learning units (LU) are representations of a paradigm for representing scientific knowledge. They were designed according to the guidelines and principles of knowledge management and knowledge engineering. Computational methods identify declarative, procedural, and structural knowledge formalisms of representation. Thus, to fully represent knowledge, the LUs paradigm includes ways to represent these three types of knowledge.

Scientific knowledge has specific facets that need to be considered for representation. Elements of declarative knowledge include the setting of the domain, the problem, the methods, and their specifications. Elements of procedural knowledge are embedded in a temporal process that includes motivations, assumptions, hypotheses, experimental design, refinements, results, and conclusions. Elements of structural knowledge are represented in how different learning units relate to each other.

We identified three categories of LUs: Background, Progress, and Complete. To capture scientific knowledge all categories have four core fields (additional fields are utilized to capture data such as date, author, etc.). While some fields are differently labeled across the LU types, they capture the same component of knowledge, just at different stages of the research process. Table 1 presents the four core fields along with (simplified) example data. The three categories of LUs follow the research process from the preliminary study of the literature to the specification of experiments to the actualized research contribution. We now describe in detail each of the unit categories and the associations.

Table 1. Complete learning unit representation

Field	Description	Example
Research Activity	The high level task that that is the focus of the learning unit	Investigating Dispsersion
Domain Contexts	Keywords specific to this particular research	Indoor Air, Anthrax, Respiration
Contribution	What was learned from this research	The deposition of aerosols on surfaces are influenced by the surface characteristics and particle diameter.
Support	What evidence supports the claims made in the contribution	Markov model of particle fate and transport

Identically named in all three types of units, Research Activity and Contexts describe the applicable task and context of the contribution. The research activity is very general to contemplate every possible variation of this same activity. Specificity is given by the domain contexts that discriminate units from each other. Both research activity and domain contexts are important elements of declarative knowledge. The Support and Contribution fields have slighted different names across the units, suited to the stage of research, but capture the same type of knowledge.

Background LUs (Background) represent scientific knowledge that becomes relevant to a user because it motivates or provides background to current research. The contribution is knowledge that was learned by accessing the literature. This contribution is validated by the citations that support it. It may list as many citations as necessary to support the statement in the contribution.

Things that are in Progress LUs (Progress) were introduced to create opportunities for sharing during the research process by allowing members to share the preliminary stages of their research activities. Users describe their Expected Contribution, along with their hypotheses. Analogously, as support, users present their experimental design.

Things that have been Completed LUs (Complete) are submitted when an experimental design has produced results, that is, a contribution is available to be shared. The results provide the empirical support for the contribution.

The Complete LU is the standard knowledge artifact that describes a scientific finding. The additional types of LU, Background and Progress, complete the paradigm to better capture the research process. Figure 2 illustrates three LUs.

Learning units embed procedural knowledge because they inherently retain a problem-solution pair. The research activity and domain contexts represent the problem. The solution varies based on the type of LU. In Background units, the solution is the learned scientific fact, substantiated by a citation. In progress units, the solution is the experimental design that is to be used to address that activity. In Complete units, the scientific fact solves the research activity and is substantiated by the experimental results.

There are two elements of structural knowledge in LUs. One is subjective. It intrinsically connects the contents in the fields of a unit. A scientific fact in a Complete unit, for example, is connected to the unit's research activity and its domain contexts as a solution to them. The second element is explicit: Associations.

There are two types of associations. One stems from the associations between the three categories that inform the sequencing among these units. The other is more complex, aiming at describing relations between units. Labels of these relations are "informs", "benefits", "motivates", "uses the output of".

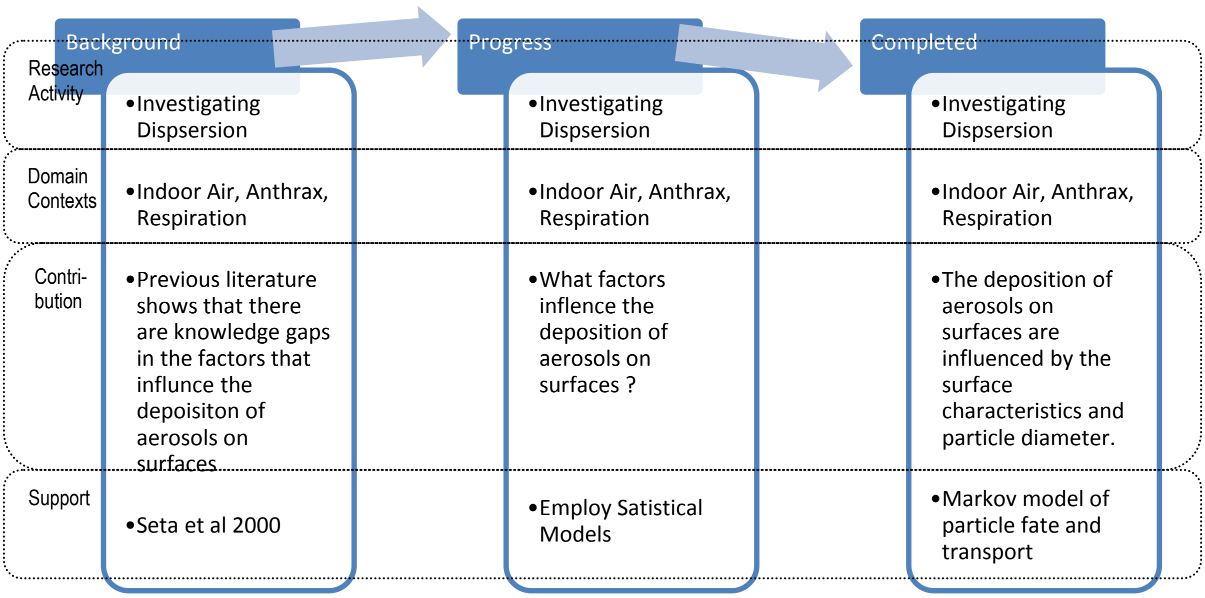

Figure 2. The three categories of learning units

4. EXAMPLE: GENERATING SLIDES FROM LEARNING UNITS

For illustration purposes, we now describe how LUs could be used to generate slide presentations. In the context of scientific presentations, a slide very often summarizes a component of research such as the underlying literature, the experiments, the findings or the conclusion (Sravanthi et al. 2009).

Any approach to generate slides applies some form of mapping from a content source into the slide design. Our approach to generate presentation contents from learning units comprises three main entities, namely, the originating learning units; the target slides; and the mapping.

Below is a plot we use for target slides. This paper does not aim at discussing the standards for slide presentations. We simply adopt widely used standards. We use the basic notion given by Alley (2003), that a slide presentation has a beginning, a series of topics, and an ending. We complete our plot with guidelines in Hofmann (2009), who expands the list of topics to three segments: Tell What You Will Tell, Tell Them (analogous to the list of topics in Alley 2003), Tell Them What You told. Our own labels become Intro, Core, and End. They are described in Figure 3. Their mappings are next.

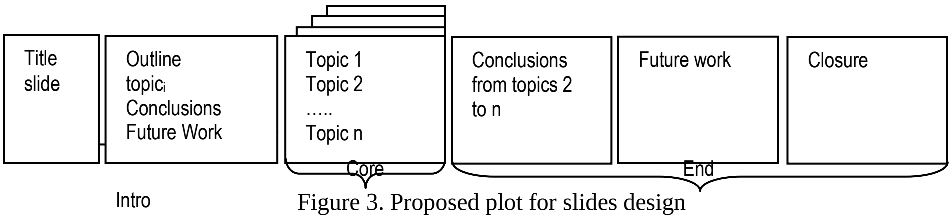

Figure 3. Proposed plot for slides design

The Intro consists of the Title and Outline slides. The Title slide is self-explanatory and can be modified by the presenter as she likes.

The Outline lists items that depend on elements from the Core and the End. The Core will produce a number of topics based on the number of LUs included in the presentation. The items in the Outline are: 1) Background; 2) one item for each topic from the Core; 3) Conclusions; 4) Future Work; 5) the Closure.

The Core consists of a list of topics corresponding to the segment Tell Them. More specifically, "Present what you studied and how you studied it. Present your results, Hoffmann (2009) pp. 519". This is the core of the presentation, which we map with LUs.

The mapping between LUs and the target slides follows parameters informed by the presenter. First, the presenter selects units of the type complete that pertain to the presentation. Each Completed unit is necessarily associated to at least one Progress unit. It is very likely that each Progress unit is associated to multiple Background units

The system returns to the presenter the list of Background units linked from the selected associated units. Once the presenter confirms the Background units as selected, the information needed for the mapping is complete. This confirmation is important to guarantee that the background, once considered pertinent, is suitable to the aspects of the completed work that is being discussed.

Slides in the End segment include Conclusions, Future Work, and The End. The conclusions will require one item per Completed unit selected for the presentation. If there are more than two or three selected units, these contents shall be divided in multiple slides. The contents come from the Contribution field of the complete units.

The Future Work slide is to include further investigations particularly one that originate from given conclusions. These contents are to be extracted from the Progress units that do not have completed units yet. To guarantee their relevance, these units should have been associated to the selected Complete units. Nevertheless, it is likely the presenter may wish to change these contents for two reasons. One is because the research tasks in the Progress units may be too specific while this slide is meant to discuss more general ideas. Other is because the presenter may not wish to reveal details of unfinished work.

The End slide may include a message of thanks, further acknowledgements, or simply the presenter's or team contact info. In particular this slide can be modified completely by the presenter.

5. CONCLUDING REMARKS

This paper introduces an approach for the semi-automated construction of learning products. It introduces the Learning Units paradigm as an intermediary machine readable representation to enable the approach.

The LUs paradigm has been shown to be useful to guide scientists to communicate scientific knowledge so it becomes computationally tractable. LUs make it possible due to its combination of declarative, procedural and structural knowledge representation elements.

One limitation of LUs is that its format favors naturalness for humans, making it less efficient for inferential adequacy and computational manipulation. This represents a trade-off between quality of elicited knowledge versus that of the final product it generates. This is why we emphasize that the approach can construct learning products semi-automatically. This may in fact be advantageous because it poses the requirement of humans in the loop, serving as an additional revision step.

Previously, we have produced report drafts automatically from learning units. In this paper, we illustrate how slides can be generated. In future work, we will explore further products like cognitive tools. We will also investigate the adaptation of one same learning product tailored to different audiences. Finally, we will continue to investigate alternative ways to capture scientific knowledge into learning units.

ACKNOWLEDGEMENTS

The authors would like to thank the ideas and references offered by the reviewers. This work is supported in part by the U.S. EPA-Science to Achieve Results Program and the U.S. Department of Homeland Security Programs, Grant # R83236201.

REFERENCES

Alley, M. 2003. *The Craft of Scientific Presentations: Critical Steps to Succeed and Critical Errors to Avoid*. Springer-Verlag, New York.

Hofmann, A. H. 2009. *Scientific Writing and Communication: Papers, Proposals, and Presentations*. Oxford University Press, New York.

Lanham, R. A. 2006. *The Economics of Attention: Style and Substance in the Age of Information*. University of Chicago Press, Chicago.

Sravanthi, M., Chowdary, C.R., and Kumar, P.S. 2009. Slidesgen: Automatic generation of presentation slides for a technical paper using summarization. *Proceedings of the Twenty-Second International FLAIRS Conference*, pp 284-289. AAAI Press, Menlo Park, CA.

van Joolingen, Wouter 1999. Cognitive tools for discovery learning. *International Journal of Artificial Intelligence in Education*, Vol. 10, pp 385-397.

Weber, R. O., Gunawardena, S. and Abraham, G. 2008. Representing and Retrieving Knowledge Artifacts. *Proceedings of the 7th International Conference in Practical Aspects of Knowledge Management*, pp 86–97. Springer-Verlag, Berlin Heidelberg.

Weber, R. O., Gunawardena, S. and Proctor, J. M. 2007. Generating Reports from Case-Based Knowledge Artifacts. *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society*, pp 445-446. AAAI Press, Menlo Park, CA.

Weber, R. O., Morelli, M. L., Atwood, M. E. and Proctor, J. M. 2006. Designing a Knowledge Management Approach for the CAMRA Community of Science. *Proceedings of the 6th International Conference in Practical Aspects of Knowledge Management*, pp 315–325. Springer-Verlag, Berlin Heidelberg.