Controlled Authoring of Biological Experiment Reports

Caroline Brun
Xerox Research Centre Europe
Caroline.Brun@xrce.xerox.com

Marc Dymetman
Xerox Research Centre Europe
Marc.Dymetman@xrce.xerox.com

Eric Fanchon
Institut de Biologie Structurale
Eric.Fanchon@ibs.fr

Stanislas Lhomme
Protein'eXpert
stanlhomme@proteinexpert.com

Abstract

We give a demonstration of an application of XRCE's controlled text authoring system MDA to biological experiment reports. This work is the result of a collaboration between XRCE's Document Content Models team, CNRS's Institut de Biologie Structurale, and Protein'eXpert, a company specialized in biotechnology based in Grenoble. We start with a brief presentation of the partners involved and their respective goals. We then give some technical background on the MDA system. Some novel features of the application are discussed, in particular how MDA can be used for integrating the formalization of an experimental protocol with its associated textual documentation.

1 Partners involved

1.1 Protein'eXpert

Whereas the human genome sequencing has now been completed, the formidable task remains of understanding the function of proteins encoded by genes. For this reason, the production of *recombinant proteins* has become an essential aspect of biomedical and biotechnology research, that is, exploratory therapeutic research, functional and structural studies. Genes are coding sequences of DNA molecules and are the templates from which proteins are synthesized.
Proteins are long linear molecules, which fold into a well-defined 3-dimensional structure. The structure of a protein determines its biological function. The synthesis of proteins from genes is performed by the complex molecular machinery present in living cells. *Recombinant DNA technology* is a set of procedures that allow the production of a protein from a given organism by another organism, which can be easily manipulated and cultured. In appropriate conditions these host cells are forced to synthesize the protein that has been artificially incorporated. Thus, by transferring a gene of interest into an organism such as *Escherichia coli* it is possible to obtain large quantities of the protein corresponding to the gene (Baneyx 99). *Each protein has a specific behavior and many parameters can vary* (Stevens 2000). Protein'eXpert[1] has developed an expertise to determine optimal production conditions for recombinant proteins, and it provides several products and services in this field. One of these services is called the *feasibility study* which is a complete and standardized protein production study including cloning, expression and solubility tests, cell fractionation, purification, refolding assay (when necessary), quality control and delivery of 1-10 mg of soluble proteins. The feasibility study has been designed to optimize protein production protocols and to give comprehensive information about protein synthesis and purification conditions. By the end of the study, proteins are delivered with a complete production protocol, an expression plasmid, and a solution proposal if the protein is difficult to express. The feasibility study is carried out by a laboratory technician under the guidance of a project manager. The technician performs all the experiments and the manager writes the final report. This study follows a complex protocol with several alternatives and potential revisions of previous steps. *The experimental part lasts for about six to ten weeks and the authoring takes several hours.*

1.2 DCM-XRCE

The DCM[2] (*Document Content Models*) team is part of the Content Analysis[3] area at XRCE[4], and explores formalisms and techniques for specifying, manipulating and exploiting the semantic structures of documents, seen as global and cohesive objects. One of the DCM projects is called MDA (Multilingual Document Authoring). MDA is an interactive system for assisting monolingual writers in the production of multilin-

[1] http://www.proteinexpert.com/
[2] http://www.xrce.xerox.com/competencies/content-analysis/dcm/
[3] http://www.xrce.xerox.com/competencies/content-analysis/
[4] http://www.xrce.xerox.com/

gual documents. This tool extends conventional syntax-driven SGML or XML editors so that semantic choices down to the level of words are possible when authoring the document content. In addition, *dependencies* between two distant parts of the document can be specified in such a way that a change in one part of the document is reflected in a change in some other part of the document (long distance dependencies).
The author's choices have *language-independent meanings* (example in the case of a drug leaflet: choosing between a *tablet* and a *syrup*), which are automatically rendered in any of the languages known to the system, along with their *grammatical consequences* on the surrounding text. Although the author is not explicitly following standards, the text produced by the system is implicitly controlled both:

Syntactically and stylistically: the choice of the standard terminology for expressing a given notion is under system control, as is the choice between grammatical variants (such as active/passive sentences) for expressing a given information;
Semantically: the consequences of a choice somewhere are reflected across the whole document, the author cannot forget to provide some information that the system requires, dependencies between semantic parameters (for instance, *pregnancy* and person *gender*) can be described.

MDA is an instance of an *interactive natural language generation* system. Early systems such as DRAFTER (Paris et al. 1995), allow the user to specify interactively an internal semantic representation, from which textual realizations can be produced automatically through a generation process. More recently, in the WYSIWYM [What You See Is What You Mean] approach, (Power and Scott 98) introduced the idea of *using the textual realization itself as the basis for interacting with and updating the internal representation.* A similar approach was adopted in GF [Grammatical Framework] (Ranta 1999-), a system which has its roots in interactive mathematical proof editors, and which provides the core model for MDA. While GF is based on higher-order constructive type theory formulation of well-formed semantic representation and has its own specific grammatical realization formalism, MDA uses a single formalism (*Definite Clause Grammars*) both for the formulation of well-formed semantic representation and of its textual realization. Both GF and MDA stress the importance of *a formal specification of the well-formedness of the semantic representation* underlying the textual realization, while (Power and Scott 98) concentrates on the formal *connections* between the semantic representation and the textual realization.[5]

The MDA home page[6] gives an overview of the capabilities and uses of the system, along with related papers, as well as a demo in the area of pharmaceutical documents.[7]

2 Aims of the collaboration

Beyond the aspects of *standardization and quality improvements of their reports*, which was a primary requirement, Protein'eXpert was interested in *producing the experiment reports more quickly*, since writing such reports is a time consuming task. Moreover, Protein'eXpert wanted to allow technicians, who run the experiments, to author at least some parts of the final reports themselves. Since MDA guides the author, this task can be given to people less experienced in writing documents without risking a decrease in quality, both at the level of the semantic dependencies to be respected, and at the level of the proper English expressions to be used (French being the commonly used language at Protein'eXpert).
From XRCE-DCM's viewpoint, the main objectives of the collaboration were to confirm the value of our previously developed methodology for describing the content and form of technical documents by working in a completely new domain, as well as to get an understanding of the potential of MDA-controlled authoring in *a previously untouched business area: experimental protocols and documentation.*
While these were the initial goals of the collaboration, an interesting and unexpected outcome of performing the concrete work gradually led us to *a novel, and more general, perspective.* We noticed the existence of a strong parallelism between the experimental protocol (what experimental steps to perform with which parameters, what decisions to take, how these decisions affect the next steps) and the structure and dependencies in the written report. It was then exciting to discover that the computational model underlying MDA was very adapted, not only to the description of the written report, *but also to the fine-grained formalization of the experimental protocol itself. In this way, we have gradually moved to a view of MDA as a convenient tool for integrating the formalization of the*

[5] This difference has several decisive theoretical and practical consequences, in particular for the connection between these systems and XML-based authoring, as well as for the definability of such notions as *life/death* of authoring choices (Dymetman 2002).
[6] http://www.xrce.xerox.com/competencies/content-analysis/dcm/mda.en.html
[7] http://www.xrce.xerox.com/competencies/content-analysis/dcm/demo/mda-demo.html

experimental protocol with its associated textual documentation.

3 The realization

3.1 Design

The first step of prototype design was to specify the structure and content of the experiment reports. With the help of the grammar writers, the biological experts produced guidelines, both at the level of semantic content and of the textual expressions to be used. It was then followed by DCM formalizing these descriptions and implementing them in the MDA formalism. Details about this formalism are given in (Dymetman et al. 2000) and (Brun et al. 2000).

During the formalization and implementation phase, XRCE used its previously developed methodology of first modeling the document macro-structure (similar to a DTD[8]), then its context-free micro-structure (what types of content choices are possible at a given point in the document), and finally the dependencies between different content elements (example: some experimental observations lead to certain obligatory choices concerning the sequel of the experiment).

To perform this formalization/implementation phase, a biological expert and a grammar developer worked *in tandem* for about 40 person-days.

A side-effect benefit of such a collaboration between biologists and computational linguists is the opportunity it offers to formally analyze the content of a set of documents to extract domain-specific knowledge: this decision leads to that result, etc.

3.2 Implementation

The generic components of the system consist in an interaction kernel, written in Prolog, connected with a Java-based GUI. The interaction kernel interprets domain-specific grammars (written in a notational variant of Definite Clause Grammars), which are used both for the specification of well-formed document content as well as for the textual realizations of this content. In the case of the reports being discussed, we developed grammars for English as well as French realization, each containing about 380 rules.

3.3 A Glance at the Interface

The following figures show some screenshots of the prototype in use. The author interacts with menus associated with underlined items and may also enter free text in dedicated boxes.

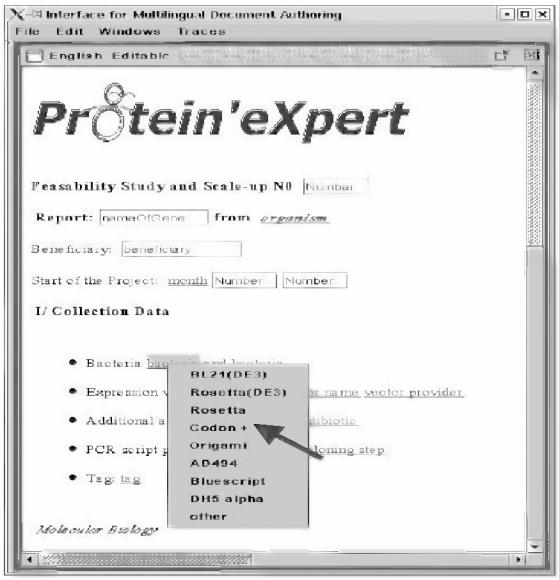

Fig.1: Interaction through menus.

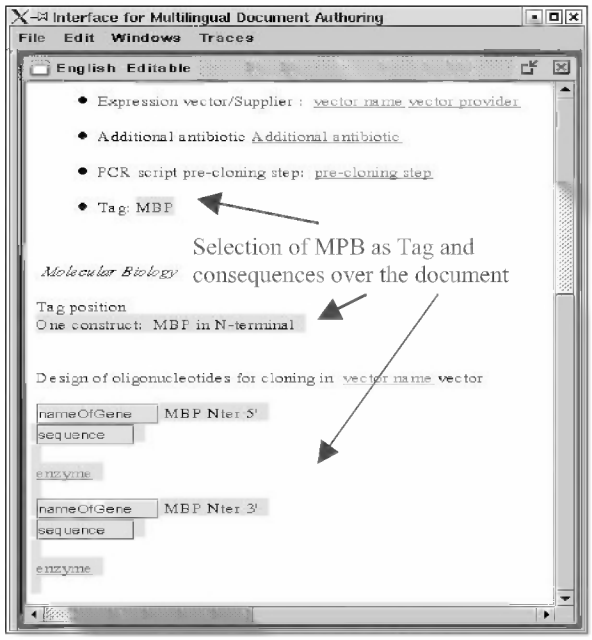

Fig.2: Consequences of a semantic choice.

4 Results

The collaboration already led to large-scale English and French grammars for the interactive authoring of biological experiments reports.

The formalization process has also been extremely valuable in inducing Protein'eXpert to be more precise in the conditions under which a certain textual expression is produced or a certain justification is given for a decision made.

[8] DTD stands for Document Type Definition

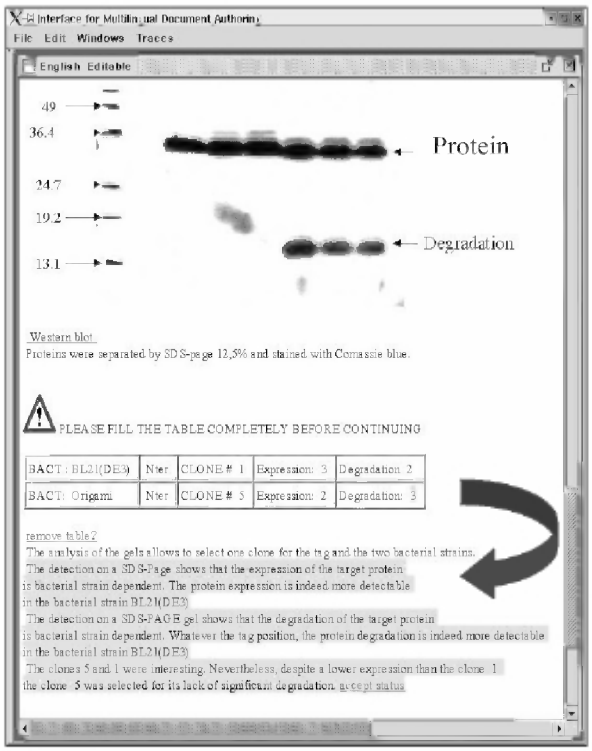

Fig.3: Analysis of a picture via a table and interactive generation of explanations.

Although this formalization process has a cost, this cost is amply repaid by the consistency and quality of the documents produced, a result that would be difficult to obtain if production of the reports where to be done manually under time-pressure.

Another interesting aspect of the collaboration is that the document class (reports on the production of recombinant proteins) has been designed without being constrained by a huge corpus of legacy documents to be accommodated. The MDA methodology is then useful, not only for producing a controlled authoring system, but as a systematic and effective way of approaching the design of new documents where a high degree of formal precision is needed.

One unforeseen and innovative outcome of the joint work has been the possibility of formalizing certain decisions taken by the biological engineers on the basis of raw experimental data. It is now possible for the author to input simply certain visual features of an image (a gel in biological terminology), and the authoring system is able to take some decisions automatically (relative to such things as a choice of bacterial strain to express the protein) and also *to provide textual justifications for these decisions (see Fig .3).*[9]

[9] The author however has the possibility of bypassing these decisions if he does not agree.

5 Evaluation and Conclusion

The prototype for experimental reports is now under evaluation *in situ* at Protein'eXpert. First results indicate that the system clearly improves the quality and speed of report production. About 30 minutes are needed for authoring a report using the system, instead of several hours previously. The *in situ* evaluation also made us discover an unexpected side of the MDA system: its didactical aspect. The system works as a self-explaining tool since the logical consequences of a given choice at a given authoring state are immediately visible to the user. Another interesting feature is that in a multi-author context (several people contributing to a given document) MDA can provide a common working frame, by allowing technicians working on different facets of the experiment to contribute to the same report.

Finally, and perhaps most interestingly, we already mentioned a new perspective opened by the current work: *MDA can be viewed as a tool for integrating formalization of the experimental protocol with its written documentation.*

The main problem identified at this point lies in the reusability and adaptability of the prototype for new classes of experiments/documents in the same domain. This is a crucial point that will be addressed in the next phases of development, in particular through work on support tools for the grammar developer.

References

Baneyx F. 1999. *Recombinant protein expression in Escherichia coli.* Curr. Opin. Biotech. 10:411-21.

Brun C., Dymetman M. and Lux V. 2000. *Document Structure and Multilingual Authoring.* In 1st International Conference on Natural Language Generation, INLG 2000, pages 24-31, Mitzpe Ramon, Israel.

Dymetman M., Lux V. and Ranta A. 2000. *XML and Multilingual Document Authoring: Convergent Trends.* In Proc. COLING'2000, pages 243-249, Saarbrücken.

Dymetman M. 2002. *Document Authoring, Knowledge Acquisition and Description Logics.* In Proc. COLING'2002, Taiwan.

Power R. and Scott D. 1998. *Multilingual authoring using feedback texts.* In Coling-ACL, pages 1053-1059, Montréal.

Ranta A. 1999— *Grammatical framework work page,* www.cs.chalmers.se/~aarne/GF/pub/work-index/index.html.

Stevens R.C. 2000. Design of high-throughput methods of protein production for structural biology. Structure 8: R177-R185.

Cecile Paris, Keith Vander Linden, Markus Fisher, Anthony Hartley, Lyn Permberton, Richard Power, and Donia Scott. 1995. A support tool for writing multilingual instructions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI),* pages 1398–1404, Montréal.