Final report of project: 2007-014-1-024

Project of Committee on Printed and Electronic Publications

Software framework for transformation of IUPAC Color Books to XML

Task Group:

Chairman:
- Valter, Bohumír

Group members:
- Jirouš, Josef
- Košata, Bedřich
- Nič, Miloslav

Objectives:

1. To establish a comprehensive plan for transforming Color Books
2. To create a software framework to support the transformation

Summary:

Google Docs based framework has been developed and it will be demonstrated at Glasgow meeting. The framework is ready for production of Color books and other materials.

Long term strategy and possibilities are discussed in the first part of the report. The plan will be used as a supporting document for Glasgow discussion.

Recommended order of book transformations based on technical criteria only:
1. Red Book
2. Blue Book
3. Purple Book
4. Silver Book
5. Orange Book
6. Green Book

Details provided in Objective 1 - section 1: Reuse of existing materials .

Objective 1: A comprehensive plan for transforming Color Books

Introduction

There are several different aspects connected with on-line production of Color books which will be discussed in following sections. The short term given in the parenthesis will be used in the following discussion to simplify discussion with the understanding that it has broader meaning that just the used term (e.g. under the term **web** is also hidden a possibility to produce CD-ROM or make an electronic version suitable for serving to mobile devices as iPhone)

1. Reuse of already existing materials in any electronic form (**Reuse**)
2. Software systems for revisions and additions of new text to the existing materials (**Authoring**)
3. Production of electronically available materials after finalization of revised/new texts (**Web**)
4. Production of printed materials from the available electronic data (**Print**)
5. A proposal of system changes (**Proposal**)

All procedures discussed in the text were tested either directly on the IUPAC materials or in other projects accomplished by Prague group in recent years.

1) Reuse of existing materials - Reuse

Used terms:

- parsing - reading any file into the memory of the computer in a form suitable for further processing
- mark-up - the way a particular display option is specified in the file (e.g. table row)
- logical information - information which can be used to infer relations between parts of text, e.g. to connect a formula with its explanation
- manual post-processing - manually editing output generated by the parsing software to correct errors and add missing information

Experience gained during the project:

A) Green book

- Most efforts was aimed at the parsing of the TEX files of the **Greenbook** with following results:
    - the book can be parsed quite successfully with the preservation of most information, but achievement of the total fidelity is impossible. The book is optimized for the final visual effects and so some logical information is irreversibly lost, especially worrying is the problem of tables which contain rows with several lines of text. The tables are arguably the most important part of the book as they contain explanations and symbols for different phenomena. In the case that the description or equation appears on several lines the TEX markup used in the book does not permit faultless automatic recognition of individual table rows

- the book is in part in the form of a free flowing text - while it is usually possible to reasonably tag individual paragraphs/notes, their relations often make sense only in the context of the whole chapter/book and so its usage outside the book is problematic
- manual post-editing of the automatically parsed text will be necessary - the TEX files does not enable 100% precision in the capture of information by the computer without human intervention - only visual appearance of the page enables decisions about logical connections
- **finalizing of the pre-parsed output will be very time consuming even with the developed software framework - a stepwise approach is recommended**

B) Red + Blue + Purple + Silver books

- The books have much less internal structure suitable for automatic parsing, it will be very non-trivial to get logical informations out of the text
- The books contain a lot of graphical data (structures, pictures, ...). We did not have in hand original authors files, just PDFs, but we expect that any automatic parsing of the files will be problematic
- While the information contained in the books will be very difficult/nearly impossible to parse automatically there are some recurring motives which can be used to simplify processing
- **developed software framework makes the transformation to web feasible on a reasonable timescale**

C) Orange book

- parts of the book are mostly terminological and so the problems are similar to ones already encountered with the production of **Gold book**
- other parts resemble the problems encountered with the **Green book**
- We did not have access to the original authoring files but judging from the available PDFs any automatic parsing will be problematic and will require a lot of manual post-processing
- **developed software framework makes their transformation to web feasible on a substantially shorter timescale than Green book but longer than on other books**

D) Gold book

- Gold book is a subject of a related project
- as it already exists in XML form it will be possible to incorporate its relevant part to products derived from other books

E) Other materials

- From the materials published by IUPAC some recommendations, e.g. GLOSSARY OF TERMS USED IN TOXICOLOGY, 2nd EDITION (Pure Appl. Chem., Vol. 79, No. 7, pp. 1153–1344, 2007.; doi:10.1351/pac200779071153) seem to be suitable for transformation to the **Web** with reasonable effort and cost using the framework

2) Software systems for revisions and additions of new text to the existing materials (Authoring)

Used terms:

- user group - any group of IUPAC people working on a single project - e.g. a subcommmittee members working on a recommendation

Experience gained during the project:

- several prototypes has been tested, several open source tools exist which can be integrated to the web pages and which enable rather convenient editing of materials by an expert - JQuery Javascript framework; TinyMCE for editing widgets; Django as server-site framework; PostgreSQL as back-end database has been extensively used in further development
- editing widgets are not suitable for committee members with average software skills without significant support from IUPAC personels in terms of consultations, software updates and trouble solvings (complications caused by different versions of Internet Explorer are especially troubling in this context)
- the documents which are currently produced by different user groups are very variable both in terms of software and used practice - their **Reuse** without extensive post-processing is very problematic
- the current processes are very individual and depend on preferences of user groups; the problems of collaborative tools, archiving, and communication are not systematically approached, but creation of a system which some people imagine (often without real experience with some real functioning system) is outside financial and personal means of IUPAC (and much bigger organizations)
- **the developed framework can be used to add new materials by people capable of editing Wiki like systems; supervison by a dedicated IUPAC personnel will be necessary in such case**

3) Production of electronically available materials (Web)

Used terms:

- AJAX techologies - communication between web servers and browsers based on XML; it enables changing just small parts of the pages without reloading the whole and effective accessing of new information and its personalised presentation

Experience gained during the project:

- in 2 years from the beginning of the project very significant development has been achieved in technologies which are used by the Prague group in IUPAC projects - Python, Django, JQuery are progressing rapidly and reaching wide audience and so their sustainability is assured
- while older electronical distributional media (e.g. CD-ROMs) are loosing their importance with the spread of high-speed internet and worldwide connectivity the importance of mobile devices is increasing; the Internet remains the most important distribution channel, but the new projects should from the beginning consider their mobile potential

- the Prague team has capacity to produced new materials based on AJAX technologies which mean with higher interactivity and flexibility
- versions of materials suitable for printing are still commonly requested (see **Print**)

4) Production of printed materials (Print)

Experience gained during the project:

- automatic printing of materials from electronic versions without manual intervention is still problematic if a reasonable print quality is expected
- reading devices as Amazon Kindle have gained a great popularity in recent months - it should be of interest to address this rapidly developing segment, especially if the zero print cost are taken into account

5) Proposal of a new System (Proposal)

Used terms:

- Web team - the team responsible for transformation of a material to the web and other electronic media
- Prague team - a Web team centered around ICT Prague (at this moment responsible for technical part of **Web** publishing of Gold book, PAC, and the IUPAC website)
- Original material - any book, publication or report available in any form which would benefit if transformed to a **Web** product
- Derived product - a **Web** product containing some or all information of the **original material** which presents the information differently to take care of possibilities and constraints connected with the **Web** publication
- Pilot product - a product from which possible usefulness can be inferred but before its finalization for general audience

5A) The print-to-web variant

1. A committee finalizes a book/material for publication in printed media (book, PAC)
2. The finalized material is send to the team responsible for **Web** publication at the same moment (as the transformation process will require months/years it will enable to publish the **Web** version as soon as possible after expiry of licensing barriers
3. A **Web team** starts with the gradual input of the material to the developed framework and produces one/several products based on the starting materia
4. When the individual product reaches a pilot stage, the **Web team** and the project committee together finalize the project
5. The process ends when the committee responsible for the original material gives its approval to publish the product as a derivative work of the original material. The committee also decides about the wording of accompanying text explaining the relation of the derived product and the original material.
6. If a new version of the **original material** is available the whole process starts with the point 1)

This variant is financed via the Project system.

5B) The web-to-print variant

1. A committee works on a book/material and reaches a stage where a useful **Web product** can be produced both as a help to chemist and as a way to elicit comments and discussions
2. A **Web team** is contacted with the information about the possibility of creation of such **Web product**. The committee and the **Web team** create an ad hoc agreement on the formats, timetable, and sharing of responsibilities during the preparation of the project

This variant is financed via the Project system.

5C) The united database variant

Introduction:

1. The concept of books is partially obsolete in the modern web environment. The books are difficult to cooperatively author and an advanced transfer of information from the book to a system which offers enhanced display capabilities is time demanding and prone to errors
2. The united database approach offers a different view on the basic unit - instead of taking a nomenclature book/material as the central piece of process it uses a particular item as the basis - equation, symbol, set of nomenclature rules to name a particular type of compound
3. The each item can be processed and discussed separately and finalized
4. Each book is still important, but not as the primary environment and rather as a collection of settled rules

Proposed databases:

- Database of physical expressions:
    - symbols
    - equations
    - constants
- Database of structures
- Database of images
- Database of nomenclature rules

There will be

- the final and public version of each database, which contains finalized and approved entries. These are guarantied never to change and always to exist, and they will receive an unique ID (a sort of DOI - they may even receive a real DOI if it is considered financially reasonable); only IUPAC selected technical personnels will have edit access to the database to implement decisions by relevant commitees
- the working version which can be edited by assigned members of individual committees. These members are supported and instructed by a IUPAC technical personnels

The database functioning is financed on long term basis. The individual derived materials from the database are financed via the Project system.

Objective 2: create a software framework to support the transformation

Google based framework has been developed and it will be demonstrated at Glasgow meeting.

The framework is ready for production of Color books and other materials.

Some comments:

- While it is possible to create an editing system on the iupac.org website and several prototypes has been tested (and a similar system works for the web site). The development of the Google docs system created an opportunity to minimize cost and take use of the Google support
- If Google stops to provide the service or starts to charge overrated fees, there is always the possibility to return back to the iupac.org server system or to use a better suited competitor
- Google services provide several formats for software downloading (and the download can be automatized to a single instruction written by a book administrator).
- The Prague team has gained extensive experience with Google system including automatic addition and maintenance of documents as it developed and uses several systems based on Google services