# $5e^{x+y}$: A Math Aware Search Engine (for CDS)

Arthur Oviedo
EPFL, Switzerland.
arthur.oviedo@alumini.epfl.ch

Nikolaos Kasioumis
CERN, Switzerland.
nikos.kasioumis@cern.ch

Karl Aberer
EPFL, Switzerland.
karl.aberer@epfl.ch

## ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings. It is an *alternate* style which produces a *tighter-looking* paper and was designed in response to concerns expressed, by authors, over page-budgets. It complements the document *Author's (Alternate) Guide to Preparing ACM SIG Proceedings Using LaTeX2$_\epsilon$ and BibTeX*. This source file has been written with the intention of being compiled under LaTeX2$_\epsilon$ and BibTeX.

The developers have tried to include every imaginable sort of "bells and whistles", such as a subtitle, footnotes on title, subtitle and authors, as well as in the text, and every optional component (e.g. Acknowledgments, Additional Authors, Appendices), not to mention examples of equations, theorems, tables and figures.

To make best use of this sample document, run it through LaTeX and BibTeX, and compare this source code with the printed output produced by the dvi file. A compiled PDF version is available on the web page to help you with the 'look and feel'.

## Categories and Subject Descriptors

X.X [**Digital Libraries**]: Miscellaneous; Y.Y [**Information Retrieval**]: Metrics

## General Terms

Theory

## Keywords

ACM proceedings, LaTeX, text tagging

## 1. INTRODUCTION


*WOODSTOCK* '97 El Paso, Texas USA

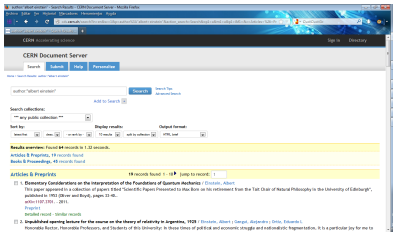

**Figure 1: CDS search result interface**

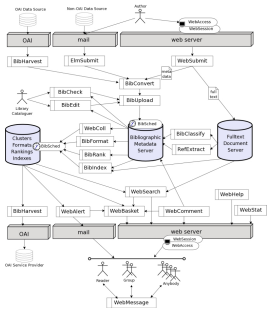

**Figure 2: Invenio global architecture**

# 2. INTRODUCTION

## 2.1 CERN

CERN (Conseil Européen pour la Recherche Nucléaire or European Organization for Nuclear Research)[**?**], founded in 1952, is the biggest research center in the area of particle physics. It is located in the French-Swiss border, near Geneva. Even though CERN directly employs around 2400 people, more than 10000 scientist from around 113 countries have visited CERN to deepen their research. As an example of its contributions to science, recently, in July of 2012, CERN announced that two different experiments (ATLAS and CMS) confirmed the existence of the particle named Higgs Boson, which lead to the award of the Nobel Prize to Peter Higgs and François Englert.

Even though the main focus of CERN relies on the study of particle physics, the strong research environment has lead to important advances in several different areas, and one of the main inventions that CERN has contributed to the world is the Web. Tim Berners-Lee, in 1989, proposed a solution[**?**] to the increasing problem of keeping track of all the information related to the different experiments held at CERN, through a hypertext system.

## 2.2 CDS & Invenio

CDS[**?**] (CERN Document Server) is the institutional digital library system developed and used at CERN[**?**]. To this date, it contains more than one million records and more than 400000 full-text documents. It provides tools to manage the complete workflow of a document taking care of the submission, annotation, editing, storage, searching, retrieval and displaying, among other phases. Figure 2.2 presents the search results for a simple query in CDS.

Invenio[**?**] is the open-source digital library software platform running behind CDS and it is a project developed in parallel to CDS at CERN. Besides CDS, Invenio supports around thirty scientific institutions worldwide including INSPIRE (a collaboration between Fermilab, CERN, DESY and SLAC) EPFL and ILO. Invenio is composed of several modules, each one of which can be mapped to a specific step in the workflow of a record. Figure 1.2 presents the global architecture of Invenio. A detailed explanation of each module can be accessed in [**?**]

## 2.3 Project overview

## 2.4 Motivation

The initial statement for this project was more generic and was encompassed as "New ways to programmatically, accurately and efficiently extract data and metadata from digital files". During a first exploration around this topic, we focused our attention to the mathematical content that is stored in these files. After more discussions we identified that the extraction, storage, indexing and finally searching for mathematical content would be a very useful project for CDS and an interesting research direction. CDS groups records in different categories or Collections such as Published Articles, Preprint, Photos, Books, General Talks among others. Some of these collections contain scientific documents in all of the different research areas where CERN is involved, and therefore a big amount of mathematical expressions are contained there. Only the Preprints collection contains 698581 records to date, harvesting documents from services like ArXiV[**?**] where most of the documents are in the areas of physics, mathematics and statistics which are very rich in mathematical content. Currently Invenio provides different ways of searching for records by specifying metadata fields like author:"Albert Einstein" or by keywords. However, there is no way to search for a given mathematical expression. The current workarounds would be to try to search for the name of the given expression if it is common enough to have been named like *Schröodinger equation*. This approach most of the times is not enough since most of the equations are not named and even named ones can be rewritten in several ways and each one may have a different importance to the user. This combination of factors, motivates the development of a complete system that allows users to search for relevant documents, based on mathematical expressions.

## 2.5 Goals

The specific goals of this project can be summarized as:

- Explore ways to automatically extract the mathematical content from a document collection
- Investigate different approaches and formats to store mathematical expressions
- Identify relevant features in a mathematical expression
- Explore ways to efficiently store and retrieve the set of features
- Implement the previous steps into a complete system (a search engine) and integrate it into the Invenio software
- Evaluate different approaches and the quality of the provided results
- Identify deficiencies in our work and propose solutions and further directions

# 3. RELATED WORK

## 3.1 Status of Mathematical Information Retrieval

Even though Information Retrieval has been a research field from around the 1950s where the first metrics for such systems were developed and in the 1960s where SMART (System for Mechanical Analysis and Retrieval of Text), the first computer based system was developed; only in the last couple years focus have been put into adapting the same techniques to mathematical content.

The Mathematics Information Retrieval Workshop[**?**] held in the context of the Conference of Intelligent Computer Mathematics in July 2012, was the first official event in the area of MIR. Here, a small competition was held to evaluate different systems and approaches.

The NII[**?**] (National Institute of Informatics) is a Japanese research institute that hosts the project named NTCIR (NII Testbeds and Community for Information access Research) which aims to build evaluation frameworks for specific subfields of information retrieval. On 2013, in the context of the 10th NTCIR conference, it hosted the NTCIR-10 Math Task [**?**] which was the first collaborative effort to evaluate different systems. The Math task was focused on two different subtasks: Math retrieval (Identifying relevant document based on given mathematical expressions and/or keywords) and Math understanding where the idea was to extract natural language information of a mathematical expression in a document. Currently submissions are open for the NTCIR-11 Math Task 2 which will be held on December 2014.

## 3.2 Mathematical based search projects

## 3.3 MIaS

MIaS[**?**] (Math Indexer and Searcher), currently, is one of the flagship projects in the indexing and searching of mathematical content. As most of the other projects, it processes documents in MathML format. It allows the user to include in the queries mathematical expressions and textual content. During indexing time, equations are transformed following a set of heuristics that include:

- Ordering of elements: Taking into account commutativity of certain operators (Addition, multiplication), elements are ordered such that $3 + a$ would be converted into $a+3$ (Since the <mi> tag comes first than the <mn> tag)
- Unification of variables: This process takes into the account the structure of the equations in despite of the naming of the variables. Expressions like $a + b^a$ and $x + y^x$ would be converted into an expressions of the form $id_1 + id_2^{id_1}$ which would match.
- Unification of constants: This step consists of replacing all occurrences of constants (<mn> tags) by a const symbol.

The extracted tokens from a given equation consist of all its valid sub-expressions. The original expression is given a weight value of 1, and following sub-expressions are given smaller weights depending on how general or specific they are. The system, also as most of the current projects, is developed by using the Apache Lucene framework. The system adapts the default scoring equation such that the given weight is taken into account. Publicly available instances of the system can be found at: `http://aura.fi.muni.cz:8085/webmias/ps?n=-1` running on the MREC[**?**] dataset and `http://aura.fi.muni.cz:8085/webmias-ntcir/` (Running on the NTCIR dataset)

## 3.4 EgoMath

EgoMath[**?**] and its new version EgoMath$^2$[**?**], is a system oriented to index the mathematical content from the Wikipedia.org database. The processing steps before indexing are similar to the ones in MIaS (Rearranging of symbols, unification of constants). The extraction process started from a complete dump of the Wikipedia database and filtering only the math articles, which are identified by the <math>" tag. This consists on around thirty thousand articles, from which 240.000 equations were identified. The processing step includes translating the equation from LaTeX into MathML format and then indexing both representation. Some additional effort is done into splitting large mathematical blocks like tables into single mathematical expressions. At the start of this work, EgoMath was publicly available at `http://egomath.projekty.ms.mff.cuni.cz/`, however at this moment the system seems to be unavailable.

## 3.5 DLMFSearch

The Digital Library for Mathematical Funtions [**?**] is a project launched by the National Institute of Standards and Technology, in 2010, as an online version of the Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables[**?**]. DLMF's main goal is to compile the mathematical knowledge in the form of equations, functions, tables and make this information useful for researchers and public in general. The system contains a search functionality that allows to input a combination of full text and LaTeXsnippets. The system tries to perform an exact match, and if no results are found, it relaxes the query. DLMF also perfoms some normalization of the equation and some cleaning of some characters, but no further details are provided. The content indexed by this project is highly curated, which differentiates it from the other projects. In the report published in 2013 [**?**], it is reported to have indexed around 38000 equations. This project is publicly available at `http://dlmf.nist.gov/`

## 3.6 MathWebSearch

MathWebSearch[**?**] (Currently on its 0.5 version) is a open-source search engine for mathematical equations. While most of the documentation relates on the architecture of the system and how they address scalability; the indexing technique is also very interesting. The system implements and idea proposed by Peter Graf in [**?**] called substitution tree indexing. This idea can be viewed as a generalization of the variable and constant unification. It represents the equations as a tree an recursively generalizes each sub-expression. A public available instance of the system is available at `http://search.mathweb.org/tema/` which runs on the Zentralblatt Math database[**?**].

## 3.7 LaTeXSearch

LaTeXSearch `http://latexsearch.com/` is a system developed by the scientific publisher Springer that allows to search over a database of around eight million latex snippets extracted from their publications. The system, unfor-

tunately, is proprietary and no further details are provided.

## 3.8 WikiMirs

WikiMirs[**?**] is a system that, as EgoMath, allows to find relevant Wikipedia entries by providing mathematical expression in LaTeX. It combines the information about the semantic tree with the layout tree of a given expression and uses a similar approach to index different generalization levels. The system is available at `http://www.icst.pku.edu.cn/cpdp/wikimirs/`

Other systems that provide similar search functionalities are Uniquation(`http://uniquation.com`) and Symbolab(`http://www.symbolab.com/`), however detailed information about the systems are not provided. Finally Wolfram Alpha[**?**] is a very powerful tool that allows to find information about lots of mathematical concepts and a wide variety of other objects (Even Pokemon).

# 4. THE *BODY* OF THE PAPER

Typically, the body of a paper is organized into a hierarchical structure, with numbered or unnumbered headings for sections, subsections, sub-subsections, and even smaller sections. The command `\section` that precedes this paragraph is part of such a hierarchy.[1] LaTeX handles the numbering and placement of these headings for you, when you use the appropriate heading commands around the titles of the headings. If you want a sub-subsection or smaller part to be unnumbered in your output, simply append an asterisk to the command name. Examples of both numbered and unnumbered headings will appear throughout the balance of this sample document.

Because the entire article is contained in the **document** environment, you can indicate the start of a new paragraph with a blank line in your input file; that is why this sentence forms a separate paragraph.

## 4.1 Type Changes and *Special* Characters

We have already seen several typeface changes in this sample. You can indicate italicized words or phrases in your text with the command `\textit`; emboldening with the command `\textbf` and typewriter-style (for instance, for computer code) with `\texttt`. But remember, you do not have to indicate typestyle changes when such changes are part of the *structural* elements of your article; for instance, the heading of this subsection will be in a sans serif[2] typeface, but that is handled by the document class file. Take care with the use of[3] the curly braces in typeface changes; they mark the beginning and end of the text that is to be in the different typeface.

You can use whatever symbols, accented characters, or non-English characters you need anywhere in your document; you can find a complete list of what is available in the *LaTeX User's Guide*[**?**].

## 4.2 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

### *4.2.1 Inline (In-text) Equations*

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin. . .\end` construction or with the short form `$. . .$`. You can use any of the symbols and structures, from $\alpha$ to $\omega$, available in LaTeX[**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

### *4.2.2 Display Equations*

A numbered display equation – one set off by vertical space from the text and centered horizontally – is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

## 4.3 Citations

Citations to articles [**?**, **?**, **?**, **?**], conference proceedings [**?**] or books [**?**, **?**] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [**?**]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide*[**?**].

This article shows only the plainest form of the citation command, using `\cite`. This is what is stipulated in the SIGS style specifications. No other citation format is endorsed or supported.

## 4.4 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material is found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed dvi output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file;

[1] This is the second footnote. It starts a series of three footnotes that add nothing informational, but just give an idea of how footnotes work and look. It is a wordy one, just so you see how a longish one plays out.

[2] A third footnote, here. Let's make this a rather short one to see how it looks.

[3] A fourth, and last, footnote.

**Table 1: Frequency of Special Characters**

| Non-English or Math | Frequency | Comments |
|---|---|---|
| Ø | 1 in 1,000 | For Swedish names |
| $\pi$ | 1 in 5 | Common in math |
| \$ | 4 in 5 | Used in business |
| $\Psi_1^2$ | 1 in 40,000 | Unexplained usage |

**Figure 3: A sample black and white graphic (.eps format).**

again, it is instructive to compare the placement of the table here with the table in the printed dvi output of this document.

## 4.5 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.eps** and **.ps** files to be displayable with LaTeX. More details on each of these is found in the *Author's Guide*.

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. and don't forget to end the environment with figure*, not figure!

Note that either **.ps** or **.eps** formats are used; use the `\epsfig` or `\psfig` commands as appropriate for the different file types.

## 4.6 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by the command `\newtheorem` and the other by the command `\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

THEOREM 1. *Let $f$ be continuous on $[a, b]$. If $G$ is an antiderivative for $f$ on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

**Figure 4: A sample black and white graphic (.eps format) that has been resized with the `epsfig` command.**

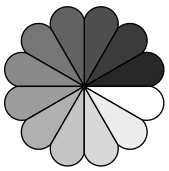

**Figure 6: A sample black and white graphic (.ps format) that has been resized with the `psfig` command.**

The other uses the **definition** environment, created by the `\newdef` command:

*Definition 1.* If $z$ is irrational, then by $e^z$ we mean the unique number which has logarithm $z$:

$$\log e^z = z$$

Two lists of constructs that use one of these forms is given in the *Author's Guidelines*.

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

PROOF. Suppose on the contrary there exists a real number $L$ such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)}\right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[**?**] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

## A *Caveat* for the TeX Expert

Because you have just been given permission to use the `\newdef` command to create a new form, you might think you can use TeX's `\def` to create a new command: *Please refrain from doing this!* Remember that your LaTeX source code is primarily intended to create camera-ready copy, but may be converted to other forms – e.g. HTML. If you inadvertently omit some or all of the `\def`s recompilation will be, to say the least, problematic.

# 5. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference

**Table 2: Some Typical Commands**

| Command | A Number | Comments |
|---|---|---|
| `\alignauthor` | 100 | Author alignment |
| `\numberofauthors` | 200 | Author enumeration |
| `\table` | 300 | For tables |
| `\table*` | 400 | For wider tables |

**Figure 5: A sample black and white graphic (.eps format) that needs to span two columns of text.**

to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

# 6. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

# APPENDIX

# A. HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e. the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

## A.1 Introduction

## A.2 The Body of the Paper

### *A.2.1 Type Changes and Special Characters*

### *A.2.2 Math Equations*

*Inline (In-text) Equations.*

*Display Equations.*

### *A.2.3 Citations*

### *A.2.4 Tables*

### *A.2.5 Figures*

### *A.2.6 Theorem-like Constructs*

*A Caveat for the TeX Expert*

## A.3 Conclusions

## A.4 Acknowledgments

## A.5 Additional Authors

This section is inserted by LaTeX; you do not insert it. You just add the names and information in the `\additionalauthors` command at the start of the document.

## A.6 References

Generated by bibtex from your .bib file. Run latex, then bibtex, then latex twice (to resolve references) to create the .bbl file. Insert that .bbl file into the .tex source file and comment out the command `\thebibliography`.

# B. MORE HELP FOR THE HARDY

The sig-alternate.cls file itself is chock-full of succinct and helpful comments. If you consider yourself a moderately experienced to expert user of LaTeX, you may find reading it useful but please remember not to change it.