Digital Mathematics Library
Some Technical and Organisatorial Suggestions

Ulf Rehmann

Abstract.

It is suggested to take DjVu as the basic format for the DML. This format is publicly documented, controlled by software in the public domain (GPL), compatible with all known image and text formats (public domain converters to and from those other formats exist). It is a multi-layer format allowing annotation and linking of the encoded digitized objects, coding and decoding software exists for all popular computer systems.

Further, for the worldwide DML, an infrastructure setup is suggested which operates in the spirit of the Open Source Movement, with cooperation among all nationalities. Software and services useful for and developed in the DML system should be offered publicly and freely, preferably under license conditions as being used for Open Source.

All services are understood to be available and useful not only in the world of mathematics, but for any scientific and public purposes.

1 Format

It is suggested to take DjVu as the basic format for the DML.

This is a file format together with a compression method, whose *underlying algorithms are designed for efficient storage, retrieval, and display of scanned document images on the World Wide Web. DjVu provides high compression rates by handling text and images differently, each in a highly efficient way. It provides progressivity, allowing an application to first display the text, then to display the images using a progressive buildup. It is designed to be implemented using data structures appropriate for efficient navigation within an image. It is lightweight: the decoder and data can reside in a small amount of memory, even for large images.*[1]

[1] Quoted from the Specifications of the DjVu Image Compressing Format[8]

The format is described in a public specification[8] and based on the EA IFF 85 format (Electronic Arts public domain IFF standard for Interchange File Format, released in January, 1985[9]).

As such, it is independent of any specific hardware or operating system.

The format is compatible with practically all other formats which so far are in use, in particular with those which have already been used for digitization projects in mathematics and elsewhere. For example, it is possible and easy to convert Dvi, Postscript, Pdf, Tiff, Jpeg, Gif, and others into DjVu, and conversely, with freely available tools[2].

This format has the advantage of a uniform approach, allowing to transform data which already exist in one of the classical and often proprietary electronic formats into a well established and publically documented one.

Contrary to other formats, it downloads faster, displays and renders faster, looks nicer on the screen and can smoothly be zoomed with no lengthy-rerendering.

The preview and the basic storing and conversion algorithms are encoded in publicly available C++ code, which is in the public domain as open source under the GNU General Public License GPL (see DjVuLibre[3]). The software can be compiled and used on all popular operating systems like Unix, Windows, Mac.

Standalone previewers as well as plugins for popular web browsers are available as source code and in precompiled binary format for many machines and operating systems.

DjVu is used by hundreds of academic, commercial, governmental, and non-commercial web sites around the world, see the DjVuLibre server for details.

The DjVu format is defined by a so-called "multi-layer model". This allows to store the image of the scanned document together with its OCR decoded ASCII text and with further annotations (layer two) including linking to other web documents. (This is done in such a way that each stored word in the ASCII layer knows the pixel coordinates of its counterpart in the image.)

Annotations and other parts of the ASCII layers of the documents can be edited and modified by standard editing tools and are – as well as all layers of the file structure – independent of proprietary software.

This setup allows to solve the following:

- Addition of meta data directly in the DjVu file
- searchability within the ordinary text
- linking to other (digital) articles, e.g., with review articles in MR or ZBL.

Under a suitable setup, it is possible to do even more:

[2] An appropriate and most useful tool for that is the freely available program ImageMagick, which uses mostly the public domain program 'Ghostscript' for format converting. It covers more than 60 digital formats, and therefore, the format radius for DjVu is at least given by this range. Further comments are given in the appendix.

In several pilot projects (supported by the Deutsche Forschungsgemeinschaft) at the University of Essen, a retrodigitization system based on the DjVu format has been developed which allowed not only to automatically set up the final data format, but also, to automatically recognize, in the scanned articles, the bibliographical data and its references, and to link them to the corresponding entries in MS and ZBL. The procedure is carefully described in [6].

These projects could certainly serve as a model for a general setup of the DML.

It should be emphasized that a decision to use the DjVu format is not exclusive: Institutions which already use another format and do not want to or cannot switch immediately can nevertheless participate. The compatibility of DjVu with other formats allows coexistence of other formats and also a migration in small steps: Format conversion, annotating and linking could be done in different steps. (See next section for such migration processes.)

2 Technical Organization of the DML

Since the build up of the DML is a long lasting task, it is necessary to set up an appropriate infrastructure.

It is suggested here to do this in an international cooperation modeled after and in the spirit of the Open Source Movement.[3]

Of course, localized centers are necessary to do the "heavy work" of the digitization of journal volumes, i.e., to scan those volumes and to produce the electronic images, and to set up servers which offer the digitized material.

On the other hand, it should also be possible for every single mathematician or for small groups to digitize scientific documents of their interest in order to offer it on their own in the Internet and/or to upload it to the DML system. Therefore, convenient software should be publicly available to perform this, and to make it easy to bring the documents into a form which is suitable for the DML. (The software should possibly be available as source code, since experience shows that this helps to adjust procedures in an evolutionary technological environment.)

But maybe even more should be set up:

We refer to an existing example: In the DjVuzone Server[3], a public "Any2DjVu" service is integrated, which allows anybody to submit any scanned or other digital document in order to transform it into DjVu format, including an OCR service on that document. That works pretty well for PS or PDF documents of medium size (a few hundred pages are not impossible), but also for scanned images.

In fact experiments have shown that it is, with this server, very easy to enrich or upgrade the image data produced by various existing digitization centers to the DjVu format and thereby transforming pure images into searchable and annotated

[3] Good and successful models for such worldwide cooperations are the system of TEX servers CTAN (Comprehensive TEX Archive Network)[2], the system of perl servers CPAN (Comprehensive Perl Archive Network)[1], or the famous Sourceforge Net[7]

documents.

(Insomuch as OCR is concerned, this so far works well for texts in English language, but, as the originators of that service admit, not yet very well for other languages.)

In an international and long lasting project like the DML, it may be desirable to set up several such servers, equipped with specific software to perform tasks like recognizing, linking, and annotating bibliographic data.

This could mean that DjVu-ing and OCR-ing of – say – French articles could be done by submitting those images to some French server, and similarly for other languages.

A worldwide network of servers of this type not only will help institutes, scientific groups, or individuals to move their scientific documents into the DML, it will also be useful for further format adjustments and migrations.

It is worthwhile mentioning here that all this infra-structural setup is independent of mathematics and can as well be used by any discipline.

This remark might be helpful for convincing funding agencies to support the endeavor of the DML.

3 Appendix: The Format Converter Program ImageMagick, and the GNU Image Processor GIMP

We mention two important freely available image processing programs which may be useful in the context of the DML:

1. ImageMagick™ is a robust collection of tools and libraries to read, write, and manipulate an image in many image formats (over 68 major formats) including popular formats like TIFF, JPEG, PNG, PDF, PhotoCD, and GIF. With ImageMagick one can create images dynamically, making it suitable for Web applications. One can also resize, rotate, sharpen, color reduce, or add special effects to an image and save ones completed work in the same or differing image format. Image processing operations are available from the command line, as well as through C, C++, Perl, or Java programming interfaces.

ImageMagick is copyright ImageMagick Studio LLC, a non-profit organization. ImageMagick is available for free, may be used to support both open and proprietary applications, and may be redistributed without fee.

It is available for Unix, Linux, Windows 2000, Windows 95/98, Macintosh, VMS, and OS/2.

Its underlying file system is the "Magick Image File Format (MIFF)", which is a platform-independent format for storing bitmap images. MIFF is a part of the ImageMagick toolkit of image manipulation utilities for the X Window System. ImageMagick is capable of converting many different image file formats to and from MIFF (e.g. JPEG, XPM, TIFF, etc.).

For further information see [5].

2. GIMP is the GNU Image Manipulation Program. It is a freely distributed software suitable for such tasks as photo retouching, image composition and image authoring.

It can be used as a simple paint program, an expert quality photo retouching program, an online batch processing system, a mass production image renderer, a image format converter, etc.

GIMP is extremely expandedable and extensible. It is designed to be augmented with plugins and extensions to do just about anything. The advanced scripting interface allows everything from the simplest task to the most complex image manipulation procedures to be easily scripted.

It is released under the GNU General Public License (GPL) and is written and developed under X11 on UNIX platforms. There is an OS/2 port in development. There is also a native win32 port available.

GIMP corresponds to the well known, but proprietary program "Photoshop".

Further details can be obtained at [4].

References

[1] Comprehensive Perl Archive Network `http://www.cpan.org/`

[2] Comprehensive TEX Archive Network `http://www.ctan.org/`

[3] DjVuZone.org `http://any2djvu.djvuzone.org/`

[4] The Gimp `http://www.gimp.org/`

[5] ImageMagick `http://www.imagemagick.org/`

[6] Michler, Gerhard O. (2001). How to Build a Prototype for a Distributed Digital Mathematics Archive Library. `http://www.emis.de/proceedings/MKM2001/michler.pdf`.

[7] SourceForge Net `http://sourceforge.net/`

[8] Specification of DjVu Image Compression Format, Version of 1999/04/29. `http://djvu.sourceforge.net/specs/`

[9] Standard for Interchange Format Files "EA IFF 85", January 14, 1985. `http://www.concentric.net/~Bradds/iff.html`.

Ulf Rehmann
Fakultät für Mathematik
Universität Bielefeld
Postfach 100131,
33501 Bielefeld
Germany
rehmann@mathematik.uni-bielefeld.de