Research Paper

Dpaper: An Authoring Tool for Extractable Digital Papers

Xiaoqiu Le[1†], Chenyu Mao[1,2], Yuanbiao He[1],
Changlei Fu[1] & Liyuan Xu[1]

[1]National Science Library, the Chinese Academy of Sciences, Beijing 100190, China
[2]University of the Chinese Academy of Sciences, Beijing 100049, China

Citation: Xiaoqiu Le, Chenyu Mao, Yuanbiao He, Changlei Fu & Liyuan Xu (2016). Dpaper: An Authoring Tool for Extractable Digital Papers.
Received: Aug. 7, 2015
Revised: Mar. 4, 2016
Accepted: Mar. 7, 2016

Abstract

Purpose: To develop a structured, rich media digital paper authoring tool with an object-based model that enables interactive, playable, and convertible functions.

Design/methodology/approach: We propose Dpaper to organize the content (text, data, rich media, etc.) of dissertation papers as XML and HTML5 files by means of digital objects and digital templates.

Findings: Dpaper provides a structured-paper editorial platform for the authors of PhDs to organize research materials and to generate various digital paper objects that are playable and reusable. The PhD papers are represented as Web pages and structured XML files, which are marked with semantic tags.

Research limitations: The proposed tool only provides access to a limited number of digital objects. For instance, the tool cannot create equations and graphs, and typesetting is not yet flexible compared to MS Word.

Practical implications: The Dpaper tool is designed to break through the patterns of unstructured content organization of traditional papers, and makes the paper accessible for not only reading but for exploitation as data, where the document can be extractable and reusable. As a result, Dpaper can make the digital publishing of dissertation texts more flexible and efficient, and their data more assessable.

Originality/value: The Dpaper tool solves the challenge of making a paper structured and object-based in the stage of authoring, and has practical values for semantic publishing.

Keywords Dpaper; Extractable paper; Digital object; Authoring tool

JDIS
Journal of Data and Information Science
Vol. 1 No. 1, 2016
pp 86–97
DOI: 10.20309/jdis.201607

† Corresponding author: Xiaoqiu Le (E-mail: lexq@mail.las.ac.cn).

1 Introduction

Communicating research results in most fields involves managing the content of multiple media types, large datasets or databases, and computational processing and emulation, among others, and thus makes the results of almost any research an interlinked and interactive collection of content resources. The current presentation of paper results are still accomplished largely via text or power-point files, making it difficult to demonstrate the results in most suitable media and to track the supporting data and related processes.

With the rapid development of information technology, the content organization and presentation of research papers are changing. Paper content is becoming increasingly more structured, semantically tagged, interlinked, and embedded with rich media, making a paper more playable and smart. STM Tech Trends 2014[①] described a new scenario being explored, where an academic article is designed to be playable and enriched with interactive data viewers, graphs, charts, visualizations, live equations, etc. STM Tech Trends 2015[②] also reported that the journal article has been at the center of a "hub and spoke" publishing model associated with videos, graphs and tables, and various digital artifacts. A publication in this sense is meant to be a software tool and service platform that provides a better understanding of its content, while supporting easy exploration of knowledge within it and related to it. These trends suggest a fundamental change in the authoring of academic papers and consequently in their patterns of use. Even dissertations, generally the most comprehensive form of academic papers, have become not only human-readable documents, but also (and maybe more so) executable and experimental platforms for readers to understand, validate, and reuse content created by their authors. Based on this vision and to supplement existing electronic dissertation writing tools, we present the design and development details of a proposed extractable dissertation authoring tool named Dpaper.

2 Related Work

A dissertation is usually represented as a simple document such as a PDF or Word file. In this form its content is static, non-executable, and not linked to its supporting primary sources and processing tools. The content is generally not structured and has no semantic tagging. This format thus inhibits in-depth exploration and verification of background context and supporting content.

[①] http://www.stm-assoc.org/2014_04_29_Innovations_USA_STM_Tech_Trends_2014.pdf
[②] http://beyondthe bookcast.com/from-stm-tech-trends-for-2015/

Research Paper

An extractable dissertation, on the other hand, is meant to be an interactive system, in which data can be extracted for reuse such as in computation, reinterpretation, and reanalysis, and the content can be reassembled, processed, and represented in various formats. An extractable document should have the following characteristics:

1) Be composed of digital objects of various media types, representing primary materials, processed results, and even processing modules to produce desired results;
2) Be structured and semantically tagged;
3) Be executable at various levels down to the level of each object; and
4) Contain modules for processing and presentation, so the various objects of the paper can be custom-organized to meet different needs.

An electronic thesis or dissertation (ETD) tool was recently designed as a data carrier and research gateway (Schopfel, Chaudiron, & Jacquemin, 2014), where studies on digital objects management and utilization were one of its active topics. A typical tool is OpenETD[③], which is both an independent dissertation submission system and a component used to implement an institutional repository by means of METS[④]/XML conversion. In the ProQuest/UMI system[⑤], rich media such as music, videos, spreadsheets, and so forth are submitted online as supplemental documents, where authors are required to fill in detailed descriptions of their contents.

Typical models for extractable digital papers are the modular paper model and the semantic publishing model. The modular paper model was proposed by Kircz (1998; 2002), where a paper is composed of multiple modules defined as information units with unique characteristics and representation. Datasets, images, music, videos, and so on are regarded as independent interactive objects or modules that can be connected to a fixed framework for exchange. A modular structure thus gives better user experience for reading as well as publishing.

With the semantic publishing model, Hunter (2006) put forward a new information format called the Scientific Publication Package (SPP). This tool is used to package raw data, provenance products, algorithms, software, text, context, and metadata, wherein scientists are able to capture, index, store, share, exchange, reuse, compare, and integrate scientific results. SPP is a compound digital object based on a number of scientific models and represented as a Resource Description Framework (RDF)

③ https://rucore.libraries.rutgers.edu/open/projects/openetd/index.php

④ The Metadata Encording & Transmission Standard (METS) schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. More details are available at http://www.loc.gov/standards/mets/.

⑤ http://media2.proquest.com/documents/Preparing+Your+Manuscript+for+Submission+Revised+31jul2015.pdf

package. Relations among internal digital objects in compound objects are either explicitly defined by the inference of ontology in the stage of metadata capture, or defined by the scientist in the SPP specifications. The workflow technology is emphasized as part of the research process to capture the processing chains of generating scientific data and provenance products. This tool thus allows scientists to describe and perform their experiments, or track the source of errors and defects in a repeatable, verifiable, and distributed way (Woutersen-Windhouwer & Brandsma, 2009).

In terms of author editing tools, Project BioLit (Fink & Bourne, 2007) and the Scientific Compound Object Publishing and Editing System (SCOPE) (Cheung et al. 2008) offer practical data explorations made from the perspective of semantic markup and compound digital objects, respectively. Project BioLit uses the National Library of Medicine's Document Type Definition (NLM DTD) to store standardized and machine-readable publications. The NLM DTD format enables articles to be archived as XML files, which include some semantic markup of the content and unique identifiers for the article and its objects (e.g. graphs and tables). The tool is supposed to facilitate the integration of the open literature and biological data (Fink & Bourne, 2007).

Ideally, the researchers themselves are the best persons to create the extractable digital objects, because they know the complete research processes, methods, experimental materials, data, and results. SCOPE has made some attempts to tap this potential. It is designed to enable scientists to easily author, publish, and edit scientific compound objects and to package scientific experiments or datasets and resources (Cheung et al., 2008). In this system, the digital objects can be published and exchanged individually. However, examples of SCOPE have shown that the generation of compound digitals needs the help of a semantic relation network, thus making the tool impractical.

3 System Frameworks

3.1 System Design

In the proposed Dpaper, a research paper is organized with a modular paper model that uses standardized digital paper templates to structure the content. Content and presentation are independent of each other during the authoring process. The whole paper is composed of a set of digital objects that can be tagged, executed, and assembled in a personalized way, and presented as Web pages or converted to an MS Word document. Digital objects are described with open metadata specifications (such as METS and Dublin Core), and a series of operation interfaces are defined to act as micro-services for certain complex digital objects.

3.2 System Processing Framework

Figure 1 illustrates the processing framework of Dpaper. The system consists of five modules: (1) structured documents representation, (2) rich media objects making, (3) digital documents editing, (4) Web page presentation, and (5) structured documents storage. Their functions are as follows:

- Structured documents representation: A general structured paper (dissertations or journals) template that is designed with descriptive specifications.
- Rich media objects making: Responsible for the management of digital objects including generation, interaction, encapsulation, and conversion.
- Digital documents editing: Various kinds of content (texts and objects) are created, edited, modified, and combined together as a paper content unit that is defined as a new structured object. Each object is automatically tagged with structural elements to make the paper smartly structure-aware. The various digital objects are assembled in the general paper template, in which objects can be associated with data.
- Web page presentation: The paper can be presented as Web pages to be browsed and published, or converted to an MS Word document.
- Structured documents storage: All the data and digital objects are stored as Web packages, while the structured information is stored in an XML file for digital document exchange or third-party reuse.

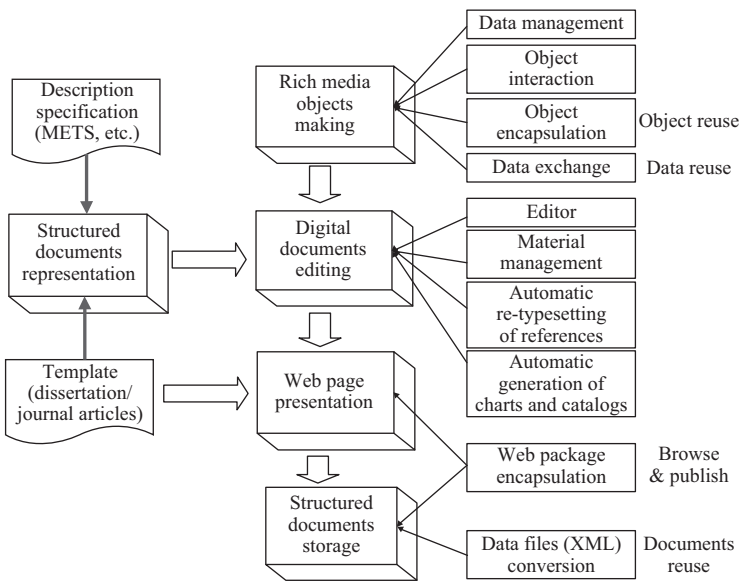


Figure 1. The processing framework of Dpaper.

4 The Key Techniques

4.1 Organization Model of Digital Objects

The content of the digital dissertation is described according to its granularity levels, where the description specifications are adopted from METS, Dublin Core, and Collection Tag Library version 3.0© designed by the National Center for Biotechnology Information (NCBI) and the National Library of Medicine (NLM).

As illustrated in Figure 2, the description framework consists of metadata objects, organizational objects and content objects. The metadata utilizes elements in Dublin Core, including title, author(s), institutions, abstract (Chinese and English), keywords, etc. The organizational objects consist of cover, declaration statements, table of contents, list of charts/figures, references, acknowledgments, appendices, and so on. The content objects include chapters and sections, and rich media objects that cover images, atlases, charts, audios, videos, animation, maps, data, network diagrams, algorithms, software, etc. Each digital object is uniformly identified by a Universally Unique Identifier (UUID). For instance, the section object that records the structure and content of a section in a chapter is described by Sec-id, Sec-title, and Sec-content. Content such as images, atlases, charts, audios, videos, and so on are embedded in the corresponding sections in the form of compound digital objects.

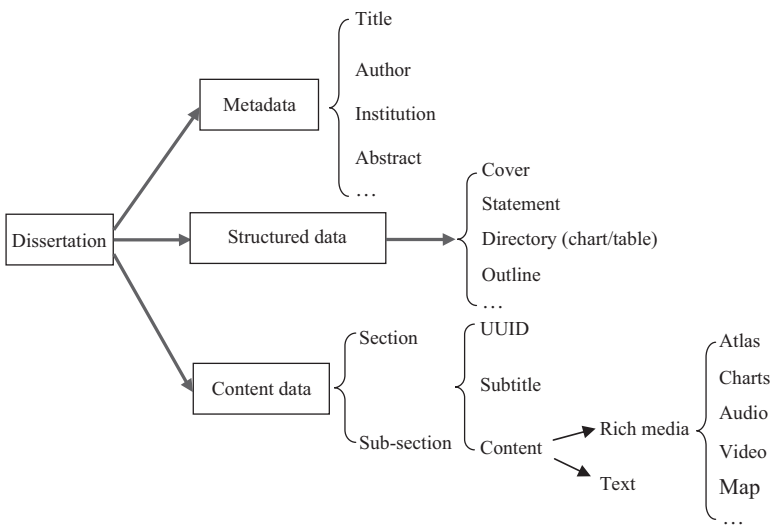


Figure 2. The organization model of Dpaper.

© http://dtd.nlm.nih.gov/book/tag-library/3.0/index.html

4.2 Communication and Interaction between Digital Objects

Compound digital objects in Dpaper integrate the entire procedures of data loading, program processing, editing, rendering, storage, etc. Data are represented as JavaScript Object Notation (JSON)[7] files that can be recalled among objects, so a data operation from an object may dynamically trigger the editing state or execute results of another object. This dynamic interaction is mainly triggered by the event mechanism, which aims to notify the digital objects to respond to the change of data when certain operations are performed.

We define several kinds of events to drive digital objects, such as Create, BeginEdit, EndEdit, Delete, Copy, Erase, and so on. The typical event response procedure is as follows:

Document operation (button/shortcut) → Environmental processing → Trigger Before events (permission checking and other related operations) → Trigger Manager events → Trigger After events → Complete events. The response process of basic events (Add Event, Modify Event, and Delete Event) in digital objects is illustrated in Figure 3.

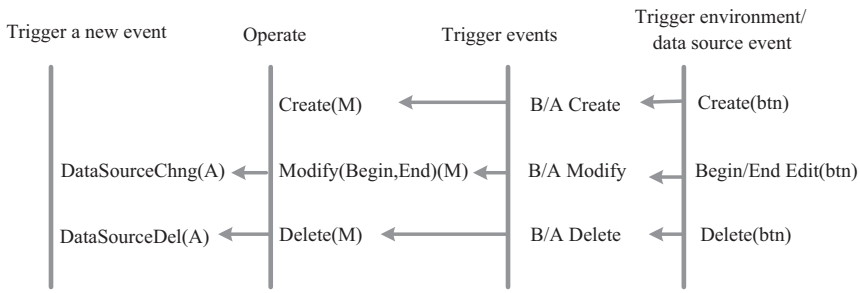


Figure 3. Response process of digital objects events.

4.3 Semantic Description of Digital Objects

With digital template technology, different digital objects defined in the digital dissertation are marked with semantic tags, and the annotation process is handled automatically by the system. To reduce the complexity of editing, two digital templates, Dissertation Template and Journal Article Template, are developed in accordance with the Master Thesis Format of the University of Chinese Academy of Sciences and with the Article Format in the *Journal of Modern Technology of Library and Information Services*. The templates are tagged as described in Section 4.1, and the format, such as font, size, style, location, and structure of the different digital objects (e.g. cover, statements, title, authors, institutions, sections, charts,

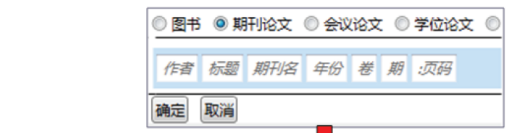
[7] It a lightweight data-interchange format (http://www.json.org/).

references, etc.) are defined specifically in templates and recorded in the system as XML documents called for editing. The semantic description examples of cover object and reference object are illustrated in Figure 4(a) and 4(b), respectively.

(a)

```
<?xml version="1.0" ?>
- <Disseration>
  - <cover>
      <Classification>G250</Classification>
      <Secret>XXX</Secret>
      <Colleage>中国科学院</Colleage>
      <ArticleType>硕士学位论文</ArticleType>
      <ArticleTitle>学术型网络数据库功能发展趋势与评价体系研究</ArticleTitle>
      <ArticleAuthor>丁艳君</ArticleAuthor>
      <GraduateTeacher>张筱</GraduateTeacher>
      <GraduateTeacherCompany>中国科学院文献情报中心</GraduateTeacherCompany>
      <DegreeType>硕士学位</DegreeType>
      <Major>图书馆学</Major>
      <ArticleTime>2009年12月</ArticleTime>
      <ArticleTimeDefence>2010年1月</ArticleTimeDefence>
      <CultivateCollege>中国科学院文献情报中心</CultivateCollege>
    </cover>
  - <coverE>
      <ArticleTitleE>The Study on Trend and Evaluation System of Resources Database Functions</ArticleTitleE>
      <ArticleTypeE>Doctoral/Master</ArticleTypeE>
      <ArticleType2E>Dissertation/Thesis</ArticleType2E>
      <MajorE>Management in the Library</MajorE>
      <ArticleAuthorE>Ding YanJun</ArticleAuthorE>
      <GraduateTeacherE>Zheng JianCheng</GraduateTeacherE>
      <ArticleTimeE>January,2010</ArticleTimeE>
    </coverE>
  - <Statement>
      <CollegeS>中国科学院文献情报中心</CollegeS>
      <AuthorS>丁艳君</AuthorS>
      <TeacherS>王企鹅鹅</TeacherS>
      <TimeS>2010.1</TimeS>
    </Statement>
  - <Abstract>
```

(b)




```
<refer_meta id="D5B536AD13B64FF0977F2028BABA954C">
  <text>[1] 周志远.<a href="http://www.cnki.com.cn/Article/CJFDTotal-
    WGYW200701009.htm" target="_blank">认知语境:关联理论对词汇语用学的解释力</a>
    [J]外国语言文学研究,2007,卷(期):页码.</text>
- <meta_data>
  - <![CDATA[
      %0 Journal Article
      %A 周志远
      %T 认知语境:关联理论对词汇语用学的解释力
      %J 外国语言文学研究
      %D 2007
      %V 卷
      %N 期
      %P :页码
    ]]>
  </meta_data>
```

Figure 4. Semantic description of (a) cover object and (b) reference object.

4.4 Reuse of Digital Objects

During the Dpaper creation, the dissertation becomes structured and partially semantic, which makes it easy for the machine to read and understand and to reuse for third-party systems. There are following three modes for reuse.

4.4.1 Reuse of Objects

Internally for a digital object, the data, programs, and library files are packaged into a separate Web package that has its interface and object description metadata, and can run without Dpaper. The copied objects can be embedded in other digital papers.

4.4.2 Reuse of Data

The data in digital objects support reuse by means of format conversion. The data in Dtable, Dcharts, Rrelation Chart, etc. can be converted into such formats as JSON, CSV[8], and RDF/XML[9].

4.4.3 Reuse of Dissertation

Internally in Dpaper, there are three kinds of XML files designed to store the structure, the data, and the formatting styles of the dissertation. The structure file records the hierarchical relationships between objects, and the style file stores the format displayed in templates, such as font, size, position, color and style. The internal structure files used for data storage and documents exchange are converted into separate METS (Figure 5).

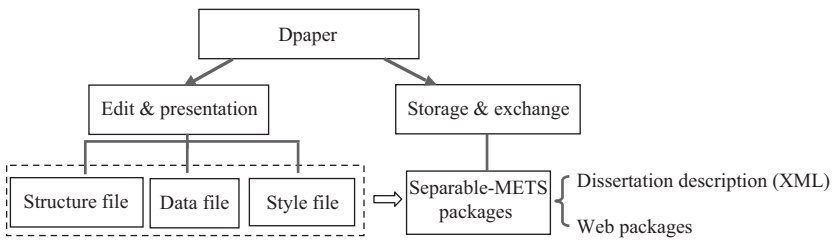


Figure 5. Storage and conversion of Dpaper.

5 Application of Dpaper

5.1 Configuration and Main Functions of the Dpaper System

A Dpaper system consists of PC client software, an MS Word plug-in, and Web application software (Figure 6). As the main platform, PC client software is designed for digital paper authoring, and includes functions of digital objects making, content editing, data management, preview, documents conversion, etc. The MS Word plug-in constructs a toolbar in MS Word for extractable digital paper authoring, and it can open digital papers created in Dpaper PC client, embed rich media objects

[8] In computing, a comma-separated values (CSV) file stores tabular data (numbers and text) in plain text.
[9] RDF/XML is a syntax defined by the W3C to express (i.e. serialize) an RDF graph as an XML document.

defined in Dpaper, and insert references automatically. So the plug-in makes it possible to author digital papers in MS Word contexts. The Web application provides the functions of data synchronization management, team collaboration, and other functions. Currently, Dpaper can run on Windows 7, Windows 8, and Windows XP, and supports Internet Explorer 8.0 or above.

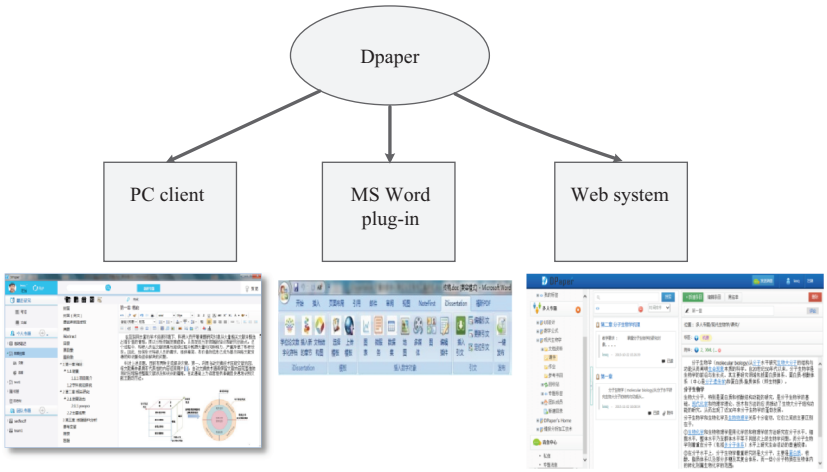


Figure 6. The tools of Dpaper.

5.2 The Development Process of Digital Papers

Figure 7 illustrates the process on how to create an extractable digital dissertation. The screenshot shows the results of a master's thesis created in Dpaper.

6 Conclusion

Dpaper explores a method to generate extractable ETDs, and presents the main ideas on digital representation, object making, presentation, and system framework. By using the digital template mechanism, a paper is represented as different types of objects, and stored as three kinds of XML files: data files, structure files, and style files, which are separated but interlinked. This kind of digital papers can be manipulated, reused, and represented as a Word document, an XML file that can be read by the machine, and runnable Web pages. Dpaper has changed an academic paper from an only human-readable document to a machine-readable document, and supports real-time playing of its objects to richly track and demonstrate the research and its data. So far, Dpaper has been released as an online version (idpaper.las.ac.cn) and currently provides a digital authoring platform for dissertations or journal papers. These papers are structured and partially semantically tagged, and can be

Research Paper

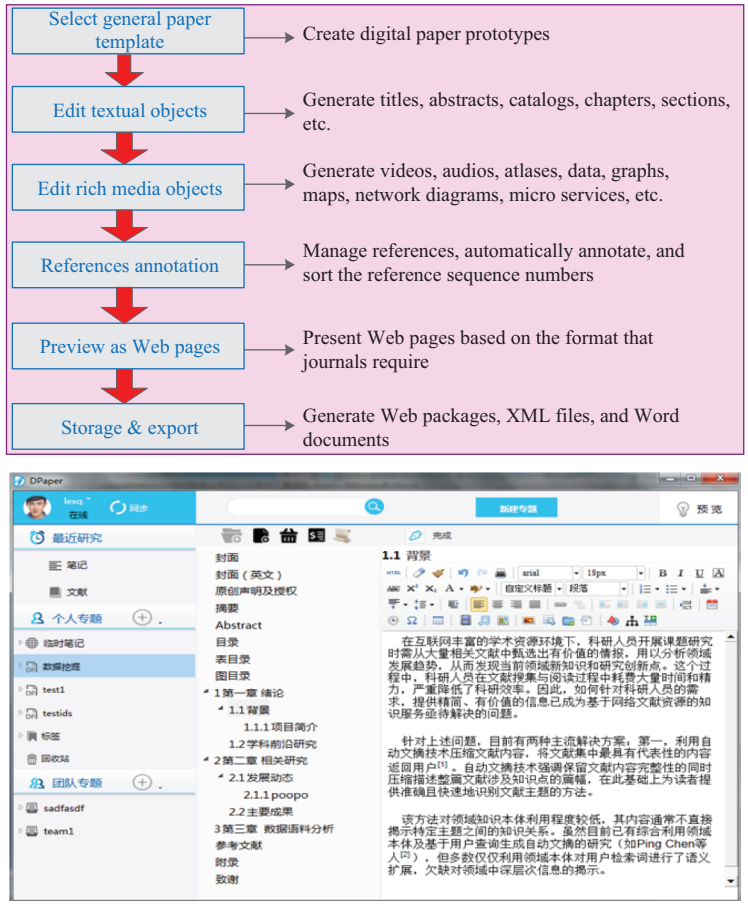


Figure 7. An example of a digital thesis.

run as a Web system. In the future, we will explore Dpaper's applicability to create other extractable academic papers such as reports and books.

Dpaper is not without limitations. Compared with MS Word, there is still a large gap in terms of general editing functions, so it needs to include objects for equations, graphics, etc. in the main platform. Also, the digital object calls between the main platform and Word plug-in need to be smoother, and its stability has to be improved.

Acknowledgements

The authors gratefully acknowledge funding support from the Chinese Academy of Sciences for the Rich Media Digital Dissertation Authoring Tools iDissertation project.

Author Contributions

Q. Le (lexq@mail.las.ac.cn, corresponding author) proposed the research idea, designed the research programs, and was responsible for system implementation. He drafted and revised the paper. Y. B. He (heyuanbiao@mail.las.ac.cn) was responsible for systems development and resolved critical issues. C. Y. Mao (maochenyu@mail.las.ac.cn), C. L. Fu (fucl@mail.las.ac.cn), and L. Y. Xu (xuly@mail.las.ac.cn) developed the modules and tested the systems. C. Y. Mao revised and edited the paper.

References

Cheung, K., Hunter, J., Lashtabeg, A., & Drennan, J. (2008). SCOPE: A scientific compound object publishing and editing system. International Journal of Digital Curation, 3(2), 4–18.

Fink, J. L., & Bourne, P. E. (2007). Reinventing scholarly communication for the electronic age. CTWatch Quarterly, 3(3), 26–31.

Hunter, J. (2006). Scientific publication packages: A selective approach to the communication and archival of scientific output. Journal of Digital Curation, 1(1), 3–16.

Kircz, J.G. (1998). Modularity: The next form of scientific information presentation? Journal of Documentation, 54(2), 210–235.

Kircz, J.G. (2002). New practices for electronic publishing 2: New forms of the scientific paper. Learned Publishing, 15(1), 27–32.

Schopfel, J., Chaudiron, S., & Jacquemin, B. (2014). Open access to research data in electronic theses and dissertations: An overview. Library Hi Tech, 32(4), 611–627.

Woutersen-Windhouwer, S., & Brandsma, R. (2009). Enhanced publications, state of the art. In M. Vernooy-Gerritsen (Ed.), Enhanced Publications: Linking Publications and Research Data in Digital Repositories (pp. 19–91). Amsterdam: Amsterdam University Press.