Wiki Authoring and Semantics of Mathematical Document Structure

Hiraku Kuroda* and Takao Namiki
Department of Mathematics, Hokkaido University,
060-0810 Sapporo, Japan

Abstract

We are developing a CMS including document authoring feature based on wiki to publish structural mathematical documents on the Web. Using this system, users can write documents including mathematical expressions written in LaTeX notation and explicitly stated characteristic structures of mathematical articles such as definitions, theorems, and proofs. Documents input to the system is published on the Web as not only XHTML files to be browsed but also XML files complying with NLM-DTD, which is used to exchange articles electronically. Not only single wiki page document, users can build a document which consist of more than one pages and is described its structure semantically by the system. In order to do this, we also propose an application of OAI-ORE and RDF vocabularies to describe structures of documents consisting of several resources.

1 Introduction

Today, many documents are published on the Web. The term "documents" here includes articles of news or blog, wiki pages, journal articles, and any other web pages. These documents are published as HTML or XHTML to be browsed, or as PDF or PS file to be printed out. Sometimes one document is published as several formats.

Some documents consist of several resources. One of examples is a document including a graphic image. Here, we assume that body text of the document is written in a HTML file and the image is a JPG file. On the Web, each of them is independent resource and given unique URI. When URI of the image is put on `src` attribute of an `img` element in the HTML file, we should treat the document as not just referencing but including or embedding the image. In this case, this document is an aggregation of two resources that are HTML file of body text and JPG file of graphical image.

Sometimes documents include not only graphic images but also whole of other (more small) documents. In general, this is called *transclusion*. HTML does not have this transclusion feature in itself, but MediaWiki, for example, has *templates* feature to extract content of other wiki pages to the page [3]. In this case, the document is an aggregation of resources that are body text written in wiki markup and other documents which are indicated in the document to be included.

Furthermore, we sometimes build a document by integrating several documents. In this case, not only a large document is just split into several documents, but each of documents is independent and has their own URI, and they can be referenced directly. In general, parts, chapters, or sections of a document are able to be independent documents. MathML specification by W3C [8] is an example of such documents. This document consists of one overview, eight sections, and eleven appendices. These divisions are independent web pages and have their own URIs. Finer divisions of a document may be independent documents according to structure or characteristics of a document. For example, definitions, theorems, proofs, and expressions in mathematical documents may be independent documents.

Open Archives Initiative Object Reuse and Exchange is standards to describe and exchange aggregations of web resources [11]. In the User Guide of OAI-ORE [13], journal articles are described as

*hiraku@math.sci.hokudai.ac.jp

Listing 1: A document including a theorem written in Wiki

```
At first, we give Taylor's theorem and proof of it.

[[theorem id="taylor_theorem" title="Taylor's theorem"|
Let $f$ be a function which is defined on the interval $(a,b)$ and suppose the $n$th
derivative $f^{(n)}$ exists on $(a,b)$. Then for all $x$ and $x_0$ in $(a,b)$,

$$ R_n(x) = \frac{f^{(n)}(y)}{n!}(x-x_0)^n $$

with $y$ strictly between $x$ and $x_0$ ($y$ depends on the choice of $x$). $R_n(x)$ is the
$n$th remainder of the Taylor series for $f(x)$.
]]

(Original text of the theorem is http://planetmath.org/encyclopedia/TaylorsTheorem.html,
retrieved at 2011.05.08)
```

aggregations of representation files such as PDF or PS. In this article, on the other hand, we propose describing documents as aggregations of constituting resources and relating the documents with their representations apart from describing aggregations.

With a background like that, we are developing a content management system **Matherial**, which manages and publishes mathematical documents and other resources. One of the purposes is developing a system which assists to write documents consisting of several resources, and publishes as web pages with appropriate metadata to describe its structure and publishes as XML files complying with NLM-DTD for further reusing.

Matherial provides authoring assistant feature based on wiki engine. Users of the system can write a wiki page including chapters, sections, mathematical statements, and expressions, or they can write some of them as independent wiki pages and integrate them into one document and publish it. Relationships between documents and included resources, and between documents and wiki pages representing them, are modeled as aggregations of OAI-ORE, and they are described in XHTML representation of documents by RDFa. Matherial can output documents into not only one or more XHTML pages, but also XML files complying with NLM-DTD. Therefore, other systems supporting NLM-DTD are able to re-use documents by Matherial.

The paper is organized as follows: In section 2, we present an example of mathematical structural documents on Matherial. In section 3, we propose an application of OAI-ORE and RDF vocabulary to describe structural documents on Matherial and more generally on the Web. Finally, section 4 concludes the paper.

2 Mathematical Contents Management System

2.1 Wiki-based Authoring

One of major features of **Matherial**, developed in this study, is assistant authoring mathematical documents. This is based on so-called "Wiki Engine", so users can write documents by simple markup notation and publish them on the Web. They can put mathematical expressions written in LaTeX notation into texts, and our own **MathML library** [7] converts them to MathML [8].

The most simple type of documents created on Matherial is one consisting of a wiki page. When users need to write mathematical text structures, such as definitions, theorems, and proofs, using functional markup for them, they can expressly provide that segments have such property.

For example, Listing 1 is a document written in wiki markup including a theorem. When users write theorems in their document directly like this, URIs of theorems are hash URIs, appending

Listing 2: A document imporint other resources

```
We begin with Taylor's theorem and its proof.

[[import wiki/TaylorTheorem]]

[[import wiki/ProofOfTaylorTheorem]]

For a function $f(x)$, $f$ is Taylor expandable when $\lim_{n\to\infty}R_n(x)=0$ where $R$ is
reminder term of [[[wiki/TaylorTheorem|the theorem]], and we have bellow.

[[import wiki/TaylorExpansion]]

Even if complex function $f(z)$ is not holomorphic at a point $c$, if $f$ is holomorphic in an
 annulus around $c$, we get Laurent series bellow,
$$f(z) = \sum_{n=\infty}^{\infty}a_n(z-c)^n$$
where
$$a_n=\frac1{2\pi i}\oint_\gamma\frac{f(z)dz}{(z-c)^{n+1}}$$
and $\gamma$ is a closed curve in the annulus(fig.[[ref annulus]]).

[[figure file/AnnulusOfLaurent id=annulus]]

This is extension of [[[TaylorExpansion]]] for functions which are not holomorphic.
```

their IDs as fragment to URI of the document. In this case, assuming a URI of a document is `http://mw2011.matherial.org/wiki/Taylor`, a URI of a theorem itself in the document is `http://mw2011.matherial.org/wiki/Taylor#taylor_theorem`.

For important definitions, theorems, and proofs, considering we discuss about them or reuse them from other documents, they should be independent documents and referenced by their own URI. On wiki of Matherial, users can set type of page, for example set that page is a theorem, the system treats the document by the wiki page as if it is described a theorem. In this case, URI of the theorem is URI of the document by wiki page. Detail about URIs and relationships of documents and wiki pages in Matherial are illustrated in section 3.

With Matherial, users can write documents importing and extracting mathematical statements which have been created as independent document. Moreover users can put images which are managed in Matherial into documents in the same way, and they can use descriptions of images which were input when images were upload to the system instead of writing new descriptions in the page. Listing 2 is a document importing statements which are already published and going on to describe a statement following them. In that example, an image referenced in the text will be imported with its description.

Moreover, aggregating these documents as sections, chapters, or parts, users can build a new document. In Matherial, users input enumeration of sub documents with metadata of the document such as title, author's information, and so on into form to build the document. Detail of semantic structure of documents which consist of several resources is described at section 3.

2.2 Output Documents

Matherial output documents as XML files. XML schemas of output XML files are XHTML [18] to read directly by web browsers, and NLM-DTD [9] to exchange articles electrically.

For XHTML files, metadata are described as RDF graph [14] and embedded by RDFa [15]. Metadata written into XHTML are metadata of the document itself such as title, authors' information, and time and date when the document was created and update, and structure information about relationships between the document and other resources. For example, relationships between resources for a document about the Laurent series shown previously is illustrated at Fig.1. The web page of this document browsed is

Listing 3: A part of an NLM-DTD XML version of a document

```
<?xml version="1.0"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD Journal Archiving and Interchange DTD v3.0 20080202//EN"
 "archivearticle3.dtd">
<article>
  <front>
    <article-meta>
      <title-group><article-title>Laurent Series</article-title></title-group>
      <contrib-group>
        <contrib>
          <name><surname>Kuroda</surname><given-names>Hiraku</given-names></name>
        </contrib>
      </contrib-group>
      <pub-date><day>29</day><month>5</month><year>2011</year></pub-date>
      <self-uri xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial
      .org/LaurentSeries/en"/>
    </article-meta>
  </front>
  <body>
    <p>We begin with Taylor's theorem and its proof.</p>
    <statement content-type="theorem">
      <title>Taylor Theorem</title>
      <p>Let <inline-formula><math xmlns="http://www.w3.org/1998/Math/MathML" display="inline
      "><mi>
*snip*
      <attrib>
        <uri xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial.
        org/LaurentSeries/en/pr/TaylorTheorem"/>
      </attrib>
    </statement>
    <statement content-type="proof">
      <title>Proof of Taylor Theorem</title>
*snip*
      <attrib>
        <uri xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial.
        org/LaurentSeries/en/pr/ProofOfTaylorTheorem"/>
        <uri xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial.
        org/TaylorTheorem" xlink:role="http://matherial.org/term/proofOf"/>
      </attrib>
    </statement>
*snip
and <inline-formula><math xmlns="http://www.w3.org/1998/Math/MathML" display="inline"><mi>&#
x3B3;</mi></math></inline-formula> is a closed curve in the annulus(fig.<xref rid="annulus">
annulus</xref>).</p>
    <fig position="float" id="annulus">
      <graphic xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial.
      org/files/2011/05/10/0/file"/>
      <attrib>
        <uri xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://mw2011.matherial.
        org/LaurentSeries/en/pr/file/2011/05/10/0"/>
      </attrib>
      <caption>
        <title>Annulus for Laurent Series</title>
        <p>Annulus for Laurent Series is shown.</p>
      </caption>
    </fig>
    <p/>
    <p>This is extension of <ext-link xmlns:xlink="http://www.w3.org/1999/xlink" ext-link-type
    ="uri" xlink:href="http://mw2011.matherial.org/TaylorExpansion">TaylorExpansion</ext-link>
     for functions which are not holomorphic.</p>
  </body>
</article>
```

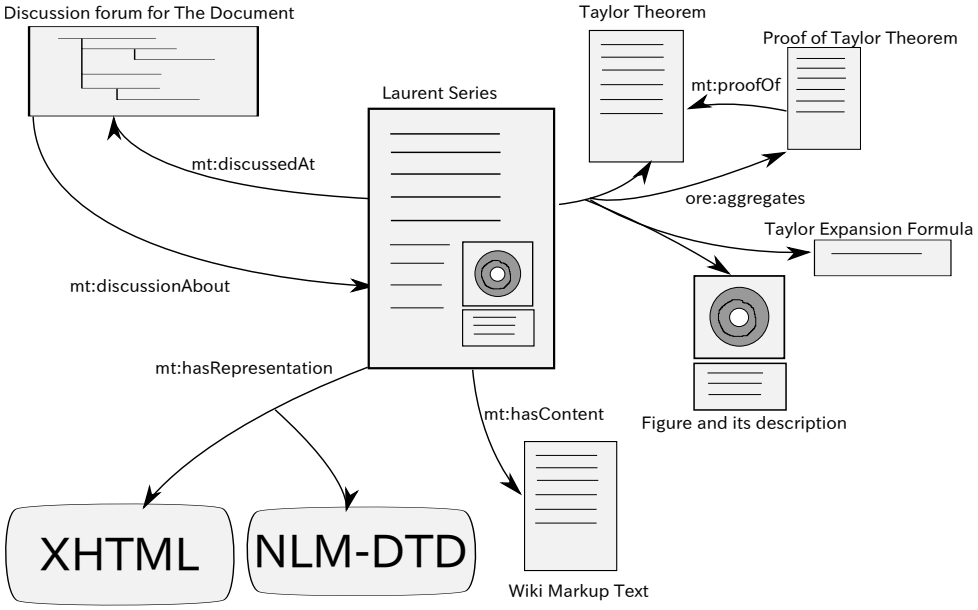

Figure 1: RDF graph of a document consisting of several resources
An RDF graph describing relationships between the document, included resources, representations and a discussion forum about the document is shown.

shown at Fig.2.

For XML files complying with Archiving and Interchange Tag Set [10] of NLM-DTD, URIs of included resources are written at `xlink:href` attribute of `uri` element in elements which included resources are extracted to. Mathematical statements are extracted into `statement` element, and type of statements are explicitly shown at `content-type` attributes. Listing 3 is a part of an XML file complying with NLM-DTD of the document shown previously. Mathematical statements are written into `statement` elements and their URLs are into `statement/attrib/uri` elements. Relationship between a theorem and its proof is shown at `statement/attrib/uri` element of the proof. Imported image and its description are extracted at `fig` element and it is referenced using `xref` element. You can get the whole of the XML file from `http://mw2011.matherial.org/LaurentSeries/en/nlm`.

3 Document structure and its metadata

The documents authoring feature of Matherial is based on wiki engine. Users write texts by wiki markup of Matherial. The most simple document is one consist of only body text but not any other resources. Matherial converts an input wiki markup text to XML files complying with XHTML and NLM-DTD, and publish them on the Web. Wiki source files, XHTML files, and XML files should be given different URI, and a URI of each files is different from the URI of the document itself (we call a URI for a document itself *platonic form URI*) [16].

In Matherial, users can write documents which include other resources, too. This does not means that documents just reference other resources. As `\includegraphics` command or `\input` command of LaTeX, users can embed images or extract contents of other documents into the document. Matherial generates XHTML files which include contents of other documents and is embedded images by `img` elements, and generates XML files of NLM-DTD which include other documents and is embedded

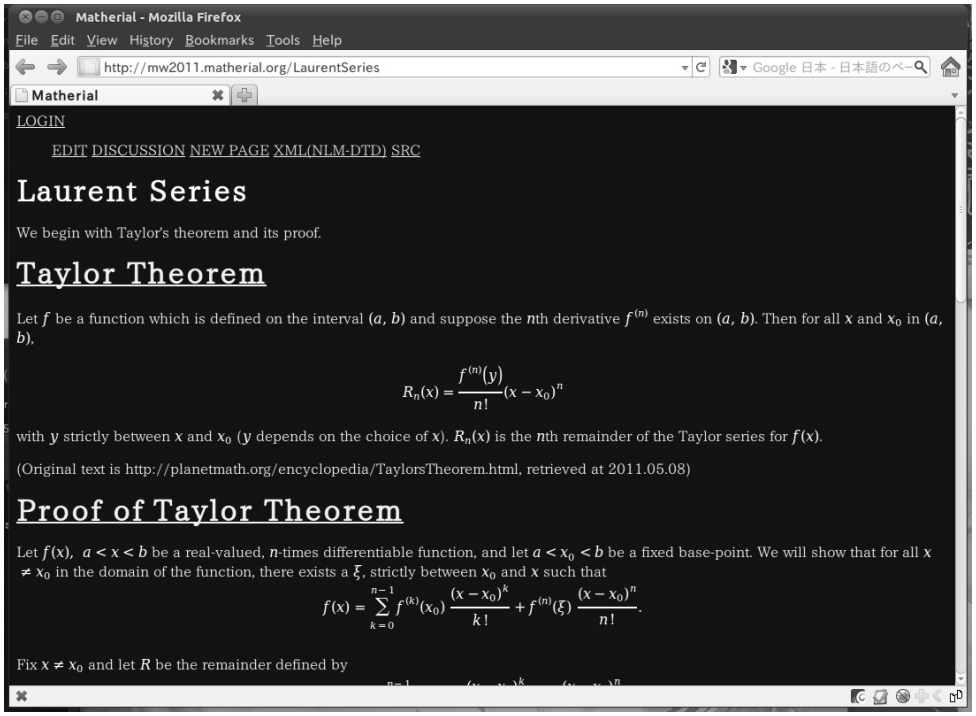

Figure 2: A web page of a document importing other documents

A web page of a document importing other documents is shown. Imported documents are extracted. Mathematical expressions written in LaTeX notation are converted to MathML and rendered by web browser. URI of this page is `http://mw2011.matherial.org/LaurentSeries/en/html`. You can also browse this page by using platonic form URI of this document, `http://mw2011.matherial.org/LaurentSeries/`.

`mt`	`http://www.matherial.org/terms/`
`rdf`	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`
`ore`	`http://www.openarchives.org/ore/terms/`

Table 1: Name spaces and Prefixes

URI of prefix `mt` is namespace for experimental vocabulary in this study. URI of prefix `rdf` is namespace for basic vocabulary of RDF [14]. URI of `ore` is namespace for vocabulary of OAI-ORE [11].

images by `graphic` elements, from wiki sources written like that.

This document structure on Matherial is described as an RDF graph whose nodes are platonic form URI of the document, URIs of resources included in the document, URIs of representations of the document, and so on (Fig.1). Matherial describes this structure by Resource map of Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [11].

The structure of a document which consists of several resources and which is represented by several representations could be applied to not only documents in Matherial but general documents on the Web. In following subsection 3.1, we will introduce OAI-ORE to describe Aggregations of resources, then in subsection 3.2, we will show a model of the document structure and metadata schema to describe the structure. RDF namespaces and its prefixes used in this article are shown at table 1.

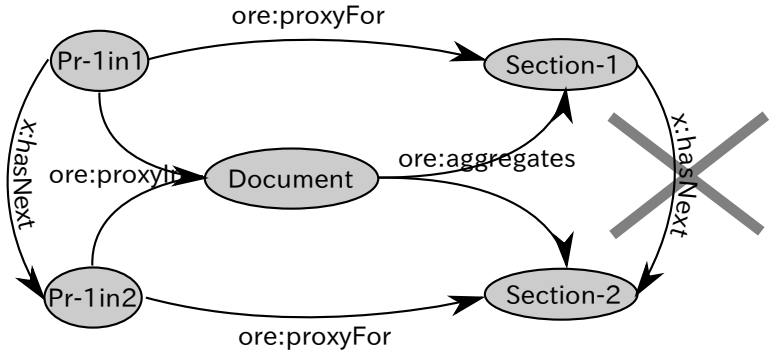

Figure 3: Proxy of OAI-ORE

`Section-1` and `Section-2` are Aggregated Resources in an aggregation `Document`. Two proxies `Pr-1in1` and `Pr-1in2` are Proxies for two Resources in the Aggregation. Order relation `x:hasNext` should not set between two Aggregated Resources directly.

3.1 Revisiting OAI-ORE

Open Archives Initiative Object Reuse and Exchange (OAI-ORE) provides a model to describe *aggregations of resources*. In OAI-ORE, a resource which consists of one or more resources is called an **Aggregation**. A resource which is a member of an Aggregation is called an **Aggregated Resource**. An aggregation is defined as a conceptual construct, so it does not have a representation. Therefore, information about an Aggregation should be described by other resources different from the Aggregation. Such a resource which describes an Aggregation is called a **Resource Map**. For example, relationships between an Aggregation and Aggregated Resources are described in a Resource Map for the Aggregation.

OAI-ORE provides a mechanism named **Proxy** to relate Aggregated Resources with properties which are given only in the Aggregation. In an Aggregation `A-1`, for example, we assume that two Aggregated Resources `AR-1` and `AR-2` have order `AR-1` $\rightarrow$ `AR-2`. Moreover, we assume that they have different order `AR-2` $\rightarrow$ `AR-1` in another Aggregation `A-2`. In this case, if we describe directly the order relationship between `AR-1` and `AR-2` in `AR-1`, it conflict to relationship in `AR-2`. This is because we describe relationship which is available only in `AR-1` independently of `AR-1`. To resolve this problem, we use **Proxy** resources which act as *Resources in the Aggregation* to describe relationships available only in the Aggregation between Aggregated Resources and other resources. For example, relationships between `AR-1` and `AR-2` in `A-1` is described using their Proxies like fig.3.

For more detail of OAI-ORE, see [11].

3.2 Structure of Document including Resources

3.2.1 Document and its Members

In this study, a **Document** is an Aggregation consisting of one or more resources. A resource which is a constituent of a Document is called a Member of Document, or simply a **Member**. An image embedded into a Document is familiar example of Member. A Member of a Document could be another Document different from the aggregating Document. In other words, aggregating several Documents, we can build a new Document. Relationships between a Document and its Members are described by RDF triples whose predicates are `ore:aggregates`.

Document Content. A Document could have a resource as one of Members which is described the content of the Document itself. We call such a Member a **Document Content**. A content of a Document is body text of the Document. When a Document includes other resources, indications to include them are written into the content. One example of Document Content is a HTML file, which is a source file of a web page. We can write not only marked-up body text, but also indications to embed images into the page using `img` elements. In this case, the Document consists of a HTML file as Document Content and images indicated. When we browse this document, web browser get a HTML file from a server, recognize it, get other resources to embed into the page, and display the completed web page. For another example, text files written in wiki markup for any wiki engines. They are described body text and embedding indications different notation from HTML. The system may convert it to a HTML, or create a PDF file which contains image files. Relationship between a Document and a Document Content is described by a RDF triple whose predicate is `mt:hasContent`, which is sub property of `ore:aggregates`.

Order of Members. When Document has the Document Content, positioning of other Members is described in the Document Content. On the other hand, when Document is simple Aggregation of resources and does not have a Document Content, we may want to describe positioning or order of Members. Furthermore, we may want to give Members complicated and non-linear order relationships such as tree structure. In this article, we propose `mt:hasNext` predicate to describe order relationships of Members in the Document. This property takes URIs of Proxies of Members for subjects and objects of triples to describe linear or complicated order relationships of Members of the Document (fig.4)

Type of Members. We may want to give Members any role in a Document. In an "article" document, for example, the first Member is abstract of the article, following some Members are Sections of the article, and the last Member is References of the article.

`mt:partType` predicate is to describe these roles of Members in a Document. This property takes URIs of Proxies of Members for subjects, and URIs of sub-classes of `mt:PartType` which represent roles of Members in Documents. Sub-classes of `mt:PartType` are `mt:Preface`, `mt:Abstract`, `mt:TableOfContents`, `mt:Part`, `mt:Chapter`, `mt:Section`, `mt:Acknowledgment`, `mt:Appendix`, `mt:References`, and `mt:Index`.

3.2.2 Mathematical Element

Mathematical documents may contain distinctive elements, such as mathematical expressions (especially *display math style*), definitions, theorems, proofs. When we write a mathematical document, preparing these elements as independent resources and including them in the Document as Members, we can reference these import elements individually and reuse them.

When we create a Document containing mathematical elements, we can show type of the Document explicitly using sub-classes of `mt:MathematicalObject`. Sub classes of `mt:MathematicalObject` are`mt:Expression`, `mt:Definition`, `mt:Theorem`, and `mt:Proof`. `mt:Theorem` has more detailed sub classes, that are `mt:Lemma`, `mt:Corollary`, and `mt:Proposition`.

A resource of type `mt:Proof` describing mathematical proof should show explicitly which theorem is proved. `mt:proofOf` is predicate for RDF triples to relate theorems and its proofs. This property takes URIs of resources of type `mt:Proof` for subjects and URIs of resources of type `mt:Theorem` or its sub classes for objects (fig.4).

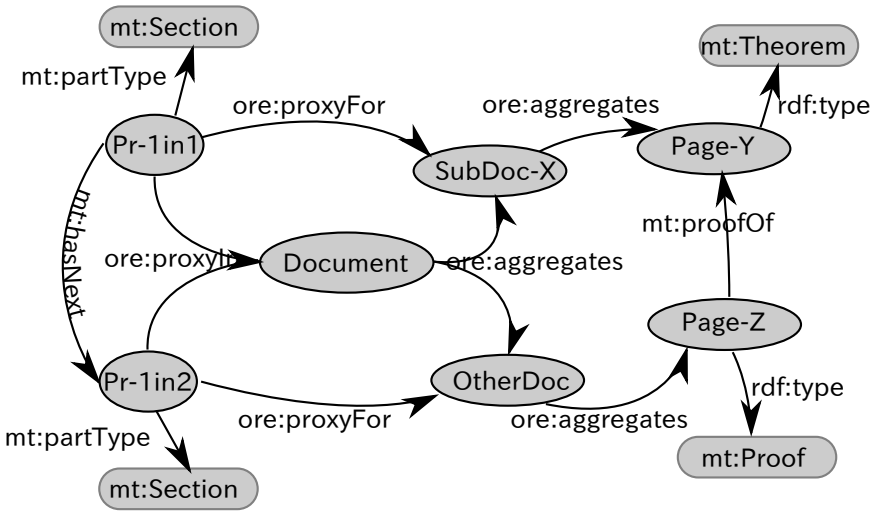

Figure 4: Structure of mathematical Document

An example of Documents which has several Documents as Members is shown. Members include independent theorem and its proof. Relationships of them are described.

3.3 Document and its Representations

In this article, Documents are Aggregations of ORE, so Documents are abstract resources and do not have entities. Therefore, to browse Documents, Documents should be related with other resources which Documents are serialized in any formats which we can browse. A resource which is serialized from a Document and has URI different from URI of the Document is called a **Representation** of the Document. When a Document is related to a Representation of the Document, We say *a Document has a Representation*. a Document Content may be one of Representations of the Document. Some Representations include all Members of Document any way like PDF files, other Representations include only indications and references to other Members like HTML files. Relationships between Documents and its Representations are described by RDF triples using `mt:hasRepresentaion` predicate (fig. 1).

4 Conclusion and Discussion

In this article, we introduced a CMS developed for authoring and publishing mathematical documents, and we proposed background semantics to describe structures of documents consisting of several resources. Using the system, we can publish not only small documents but also large documents consisting of several resources by writing in easy mark-up. Structures of documents consisting of several resources are described as RDF graphs based on Resource map of OAI-ORE, and they will be reused by Semantic Web Technologies.

Some applications of OAI-ORE are describe an article as an aggregation of OAI-ORE. The FORE-SITE [2] project developed a toolkit to describe metadata of articles from JSTOR by using OAI-ORE. In the project, each issue of journals is an Aggregation of articles, and each article is an Aggregation of individual page images and a PDF-formatted version of the entire article [1]. The ICE-TheOREM project [17] provides thesis authoring and publishing systems. In this project, each thesis is an Aggregation of sections and PDF, DOC, and ODT version of the article, and each section is an Aggregation of PDF, DOC, and ODT version of the section. These applications treats each article as an Aggregation

of parts of it and its Representations. On the other hand, in this article, we describe a Document as an Aggregation of parts of it, and we use another property for relationships between a Document and its Representations.

The OMDoc format is a content markup scheme for mathematical documents [4]. This format is designed for the Mathematical Knowledge Base. SWiM is a semantic wiki for Mathematical Knowledge Management using OMDoc and OpenMath [5][6]. While these are aimed at building the Mathematical Knowledge Base, the Matherial is aimed at publishing mathematical documents by using simple notation for authoring and only presentation markups for outputting. OMDoc also provides a document ontology [12]. RDF classes i.e. `Definition`, `Theorem` (and so on), `Proof`, and `Formula` and RDF properties i.e. `proves` and `provedBy` are defined, but a class for general mathematical expression is not defined. These classes of OMDoc ontology are subclass of `MathKnowledgeItem`. However, documents of Matherial are not expressed in OMDoc. So we use classes in `mt` namespace instead of OMDoc ontology.

References

[1] H. V. de Sompel, C. Lagoze, M. L. Nelson, S. Warner, R. Sanderson, and P. Johnston. Adding eScience Assets to the Data Web. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web*, June 2009.

[2] FORESITE. `http://foresite.cheshire3.org/`.

[3] Help:Templates MediaWiki. `http://www.mediawiki.org/wiki/Help:Templates`.

[4] M. Kohlhase. *OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]*, volume 4180 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] C. Lange. Mathematical Semantic Markup in a Wiki: the Roles of Symbols and Notations. In *Proceedings of the 3rd Semantic Wiki Workshop (SemWiki 2008) at the 5th European Semantic Web Conference (ESWC 2008)*, June 2008.

[6] C. Lange. SWiM – A Semantic Wiki for Mathematical Knowledge Management. In *The Semantic Web: Research and Applications 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings*, June 2008.

[7] MathML Library. `http://rubygems.org/gems/math_ml`.

[8] Mathematical Markup Language (MathML) Version 2.0 (Second Edition). `http://www.w3.org/TR/MathML2/`.

[9] Journal Archiving and Interchange Tag Suite. `http://dtd.nlm.nih.gov/`.

[10] Archiving and Interchange Tag Set. `http://dtd.nlm.nih.gov/archiving/`.

[11] ORE Specifications and User Guides - Table of Contents. `http://www.openarchives.org/ore/1.0/toc`.

[12] OMDoc Document Ontology. `http://kwarc.info/projects/docOnto/omdoc.html`.

[13] ORE User Guide - Primer. `http://www.openarchives.org/ore/1.0/primer`.

[14] RDF - Semantic Web Standards. `http://www.w3.org/RDF/`.

[15] RDFa in XHTML: Syntax and Processing. `http://www.w3.org/TR/rdfa-syntax/`.

[16] L. Richardson and S. Ruby. *Restful Web Services*. O'Reilly Media, 2007.

[17] P. Sefton, J. Downing, and N. Day. ICE-Theorem - End to end semantically aware eResearch infrastructure for theses. *Journal of Digital Information*, 11(1), Jan. 2010.

[18] XHTML 1.1 - Module-based XHTML - Second Edition. `http://www.w3.org/TR/xhtml11/`.