A Metadata Model Based on the Concept of Structured Digital Object(SDO) and Its Application in Digital Libraries
-From Concept to Prototype System

Ying LI, Hidehiro ISHIZUKA
liyingli@slis.tsukuba.ac.jp
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan

Abstract: **Metadata is data about data. It is considered an ideal and a very useful solution in describing/managing resources on the Internet. Especially, in the digital library field, metadata plays an important role for integrating digital resources and offering information service. However, from users standpoint, information service offered nowadays by any digital library is unable to satisfy their diversified needs, such as, knowledge information, individualization information, reusable and sharable information. A metadata model based on a new concept is required.**
As an extension of our previous research issue, we proposed the concept of Structured Digital Object (SDO is an abbreviation). SDOs are used for reorganizing/restructuring existing digital resources, because SDO set is structured data about data, this paper call it metadata. Using the metadata model based-concept of SDO, restructure various existing resources in existing digital libraries, form the so-called "Global Digital Library ". The Global Digital Library can adopt Web Services for information services. It can solve not only interoperability problems among heterogeneous resources, heterogeneous systems and operating systems, but also can meet individual user need to different granularities information. We also used Topic Maps to associate SDOs with information resources that can be located in existing digital libraries or the global digital library. Furthermore, because all the metadata is described by XML, achieve reusing, sharing information.
In the paper, we give some demonstrations to approve the points described above.

Keywords: **Structured Digital Object(SDO), Metadata Model, Global Digital Library, Web Services, Topic Maps.**

1 Introduction[1,2,3]

Metadata can be defined literally as "data about data," but the term is normally understood to mean structured data about digital resources. It provides the underlying foundation upon wihch digital resources management systems rely to provide precise access to relevant resources. Metadata required to describe the highly heterogeneous, multimedia objects on the Internet is infinitely more complex than simple metadata for resource discovery of textual docments throught a library database. In digital resources, there exists a wide variety of metadata formats, through relatively simple formats like the Dublin Core Metadata Element Set (DCMES) and the more detailed Text Encoding Initiative (TEI) and MARC, to highly specific formats like the FGDC (The Federal Geographic Data Committee). Among these, the one of the most widely known international metadata standard is the Dublin Core, an initiative that has a deliberate focus on simple resource discovery.

Most current discussions of metadata in the library and information communities have centered on issues of resource description and discovery. Creation and use of metadata is likely to become an important part of all the digital libraries.

However, some of the major disadvantages of metadata were unreliability, subjectivity, and lack of interoperability with respect to syntax, semantics, vocabularies, languages,

and underlying models. Consequently, there are many researchers currently investigating strategies to overcome different aspects of these limitations in an effort to provide more efficeent means of organizing resources in digital libraries.
The goal of this article is to propose a metadata model based on the concept of Structured Digital Object (SDO)that is expected to improve ability to access to relevant resources in digital libraries regardless of the domain or format. Using the metadata model based-concept of SDO, restructure various existing resources in existing digital libraries, form the "Global Digital Library ". The Global Digital Library is a portal. It can satisfy users for precise search, meet individual user need to different granularities information. It can also provide Web Services for information services, and the function of Topic Maps to associate SDOs with information resources. By the system functions, we can solve interoperability problems, and achieve reusing, sharing information.

2 The Key Relevant Research Areas

In this section, we have identified waht we consider to be some the key metadata research areas, both now and over the next few years. For each of the research areas, we give a brief description of the work being undertaken and some key citations.

- **Semantic Web**: “The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and peple to work in cooperation” (Berner-Lee, Hendler, &Lassila, 2001). There are two main building blocks for the semantic Web :
  - Formal Languages-RDF, DAML+OIL, and OWL (Web Ontology Language), which is being developed by the Web Ontology Working Group of the W3C.
  - Ontologies-communities will use the formal language to define both domian-specific ontologies and top-level ontologies to enable relationships betwenn ontologies to be determined for cross-domain searching, exchange, and information.
- **Web Services** (W3C Web Services Activity, 2003) are a relatively new concept, expected to evolve rapidly over the next few years. They could be the first major practical manifestation of Semantic Web-based thinking. Detailed-definition vary, but Web services will enalbe the building of software applications without having to know who the users are, where they are, or anything else about them. In the next few years, Web services may be develped that can be understood and used automatically by the computing devices of users and of public libraries. External Application Services Providers (ASPs) may also provide such services. Web services are based on open, Internet standards. The core standards and protocols for Web services are being developed and are expected to be finalized by 2003. They include (in addition to XML)
  - Web Services Description Language (WSDL)(WSDL,2003), which enables a common description of Web Services;
  - Universal Description, Discovery, and Integration (UDDI)(OASIS,2003) registries, which expose information about a business of other entity and its technical interfaces.
  - Simple Object Access Protocol (SOAP)/XML Protocol(W3C XML Protocol Working Group, 2003),Which enables structured message

exchange between computer programs

- **Metadata Harvesting**: The Open Archives Initiative (OAI) provides a protocol for data providers to make their metadata and content accessible-enabling value-added search and retrieval services to be built on top of harvested metadata.
- **Automatic metadata exraction**: technologies to enable the automatic classification and segmentationof digital resources. In particular, automatic image processing, speech recognition, and video-segmentation tools will enable content-based querying and retrieval of audiovisual content.
- **Search engines**:
  - Smarter agent-based search engines;
  - Federated search engines;
  - Multimedia searh engines;
  - Multilingual search engines;
  - New searh interfaces-search interfaces that present results graphically;
  - Automatic/dynamic aggregation and generation of search results into hypermedia presentation.
- **Topic Maps** (Topic Maps, 2000) is a new ISO standard for system describing knowledge structure and associating them with information resources. Topic Maps create a "virtual map"above the resoreces, leaving them unchanged.
- **Annotation Systems** The motivation behind annotation systems is related to the issue of metadata trust and authentication-users can attach their own metadata, views, opinions, and recommendations to particular resources or documents on the Web, which can be read and shared with others. The basic philosophy is that we are more likely to value and trust the opinions of people we respect than metadata of unknown origin.
- **Metadata schemas** for resource description such as Dublin Core, IMS (Internet Global Learning Consortium) and LOM (IEEE Learning Object Metadata) and domian specific markup languages such as MatML (Materials Markup Language), MathML (Mathematical Markup Language) or CML (Chemical Markup Language) have evolved dramatically during the last five years, Much of this development, however, has been a parallel evolution. There is often no clear indication of whether or how resource level metadata should be integrated most effectively with domain-specific content markup or with structural markup meant to describe the internal architecture of resources.

Each of the key metadata research areas mentioned above alone cannot describe the rich, granular, associative and recombinant digital objects potentially contained in digital libraries. Therefore, powerful mechanisms for content and structure description of digital libraries are required.

3 A Metadata Model Based on the Concept of Structured Digital Object (SDO)[4]

In our study, "***Metadata is a set of structured digital objects about resources in digital libraries.***" We use the definition throughout this paper.

To be faced with a great number of existing digital libraries and not to be able to precise Information Architecture is the discipline dealing with the modern version of this

problem: how to organize existing digital libraries so that users actually can find what they are looking for? Now metadata is creating as another information architecture. Metadata is the foundation of all information retrieval. It is generally assumed when organizing information that it consists of discrete objects. We use the term *object* here for the entities being organized. In computer science metadata typically means schema information, administrative information, and so on. However, in content management and information architecture, metadata generally means "information about objects", that is, information about a document, an image, a reusable content module, and so on. Since it is the management of content we are primarily concerned with here, we gave the definition, “metadata is a set of structured digital objects about resources in digital libraries.”

Our study on Structured Digital Object was started from 2002. In the earliest stage of development [5], to satisfy researchers’ high-precision search needs, we proposed the concept of the “Computer-processable Digital Object (CPDO is an abbreviation)” as retrieval target , and introduced hierarchy structure to describe granularity of CPDOs. Granularity refers to how finely you chop your metadata. For example, in the standard for encoding the full text of books using the Text Encoding Initiative (TEI) schema, a book author may be recorded as <docAuthor> ISHIZUKA, Hidehiro</docAuthor>. That’s all well and good, if you never need to know which string of text comprises the author’s last name and which the first. Because we chop metadata into elementary (atomic) units, different granularity of scientific information can be described/expressed by combining CPDOs. It makes it possible to meet all the researchers’ needs for different granularity by combining CPDOs. Meanwhile we gave some concrete examples to illustrate how transform existing different resources into CPDOs, and reasoned that it is improved precision to transform existing resources in digital library into structured information based on CPDOs.

In the second-stage [6], concerning implementation, we reified the concept of CPDO and proposed the concept of “Computer-Processable Structure Digital Object based on XML(CPSDO/XML is an abbreviation)”, which is composed of elementary units of information, and designed an information retrieval system based on CPSDO/XML for Web resources, and developed a prototype using Merck Index data.

As an extension of our previous study, In this paper, we proposed a metadata model based on the concept of Structured Digital Object (SDO is an abbreviation). The Metadata set of SDOs is resource level metadata. Using the metadata model based-concept of SDO, restructure various existing resources in digital libraries, form the "Global Digital Library ". It is a portal, and can satisfy users’ high-precision search. Via the Global Digital Library, you can use Web Services for information services. Up until now, to express relationships among SDOs, we adopted hyperlink. For improvement, in this paper, we introduced Topic Maps, and found it is effective, because it is a network structure metadata and help us to navigate cross the Global Digital Library and existing resources in digital libraries.

4 Architecture of the Prototype/Design for a Global Digital Library based on SDO Metadata

Basically, there is no one-size-fits-all architecture that could address all problems in the

digital library field. Our proposed architecture focuses on granular, associative, reuse digital object and Web Services and Topic Maps technologies.
The architecture comprises two parts as depicted in Figure 1: Creating Metadata and accessing/replying. Details are shown in Figture 2 and Figure 3.

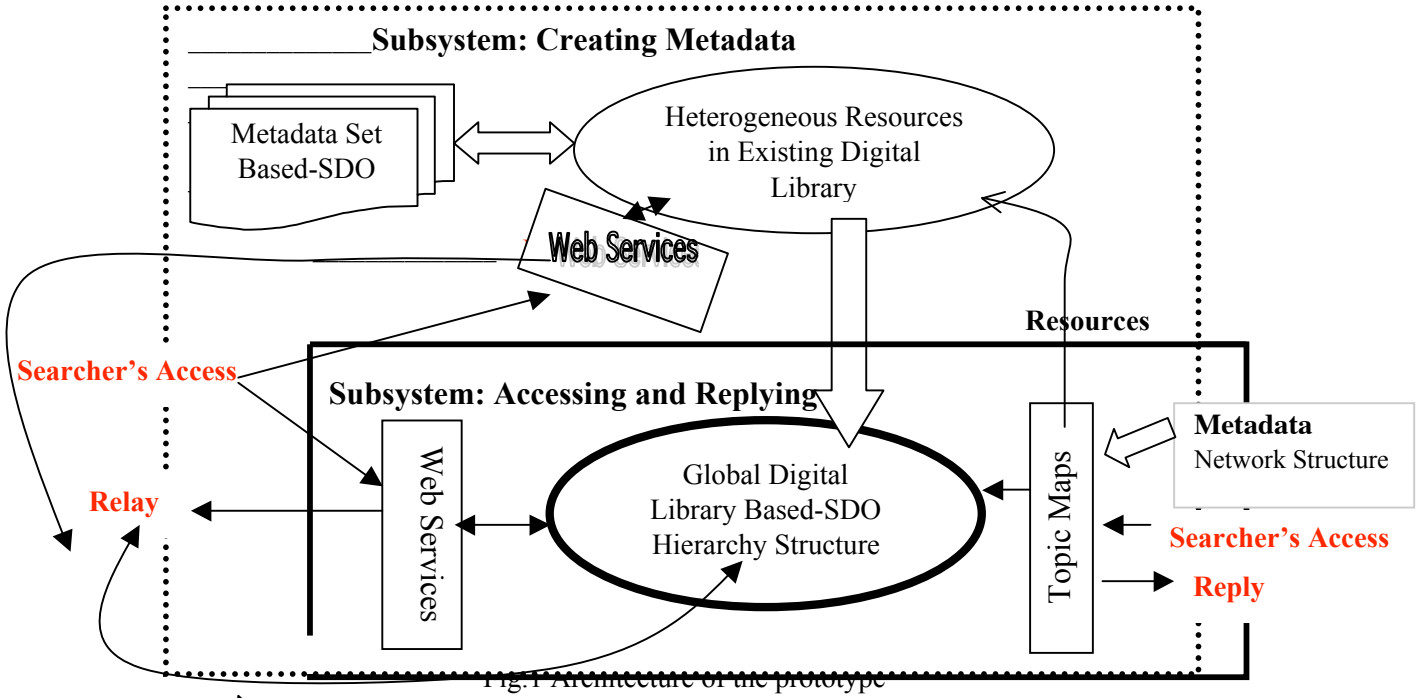


Fig.1 Architecture of the prototype

Means Matching and Forming

Means Data flow for accessing and replaying

- Web Services as search interfaces could be accessed or provided by Global Digital Library Based-SDO, as well as by any existing digital library.
- Adopting Topic Maps not only help us to navigate cross the Global Digital Library and existing resources in digital libraries, but also could implement relationships of SDOs.

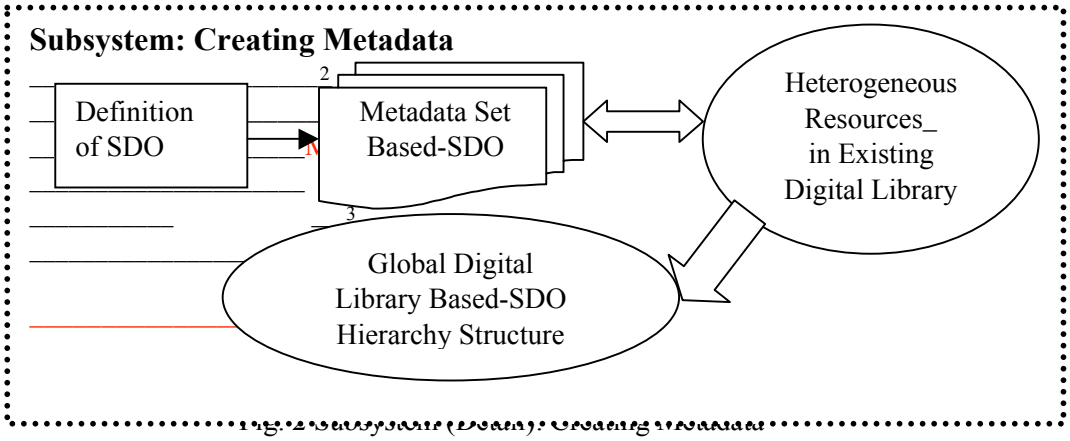


Fig.2 Subsystem (Part1): Creating Metadata

1Organizing and creating a metadata set based-SDO, which is hierarchy structure
2Matching Metadata Set based-SDO and digital objects belong to a resource
3Forming the global digital library based-SDO

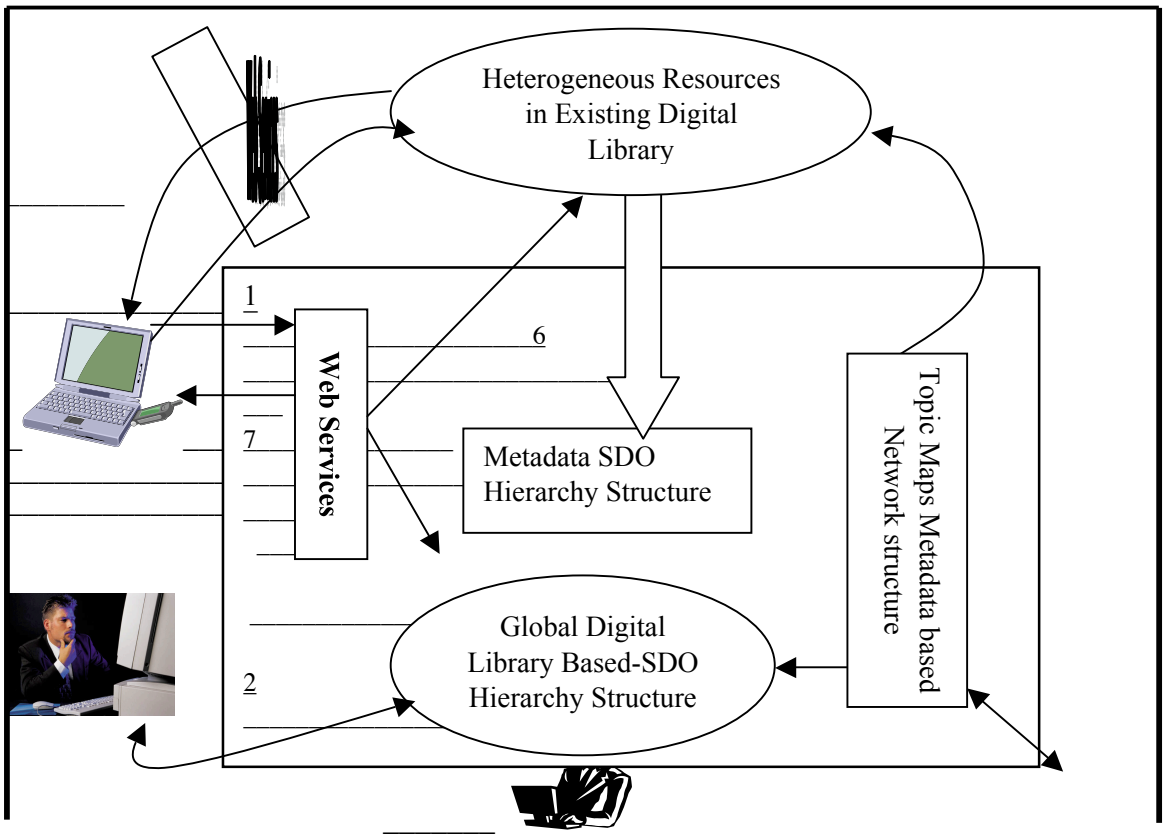
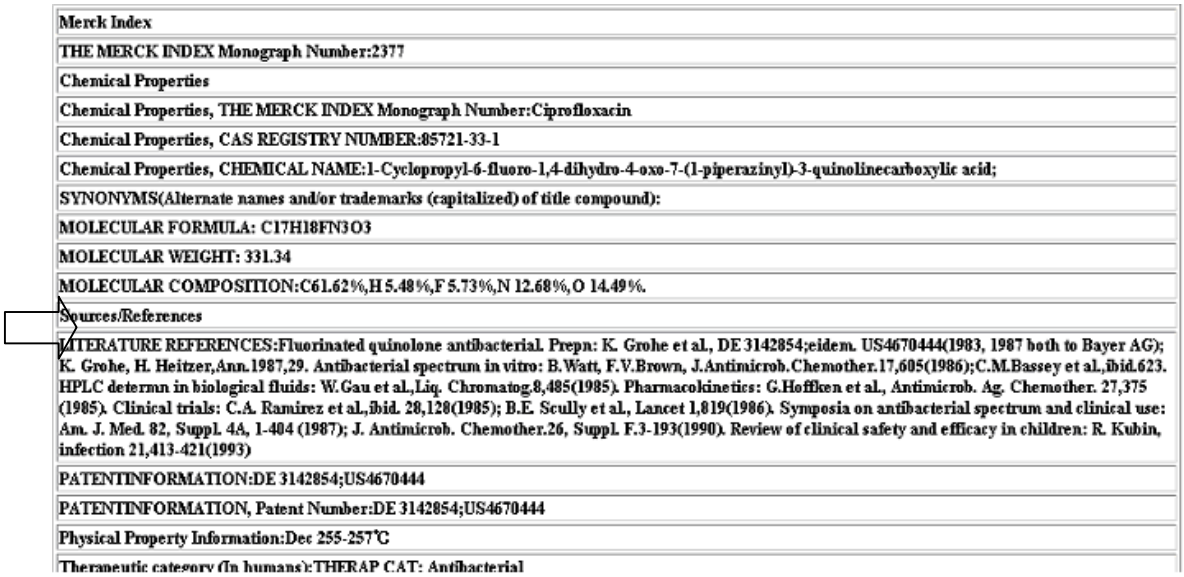


Fig.3 Subsystem (Detail): Accessing and Replying

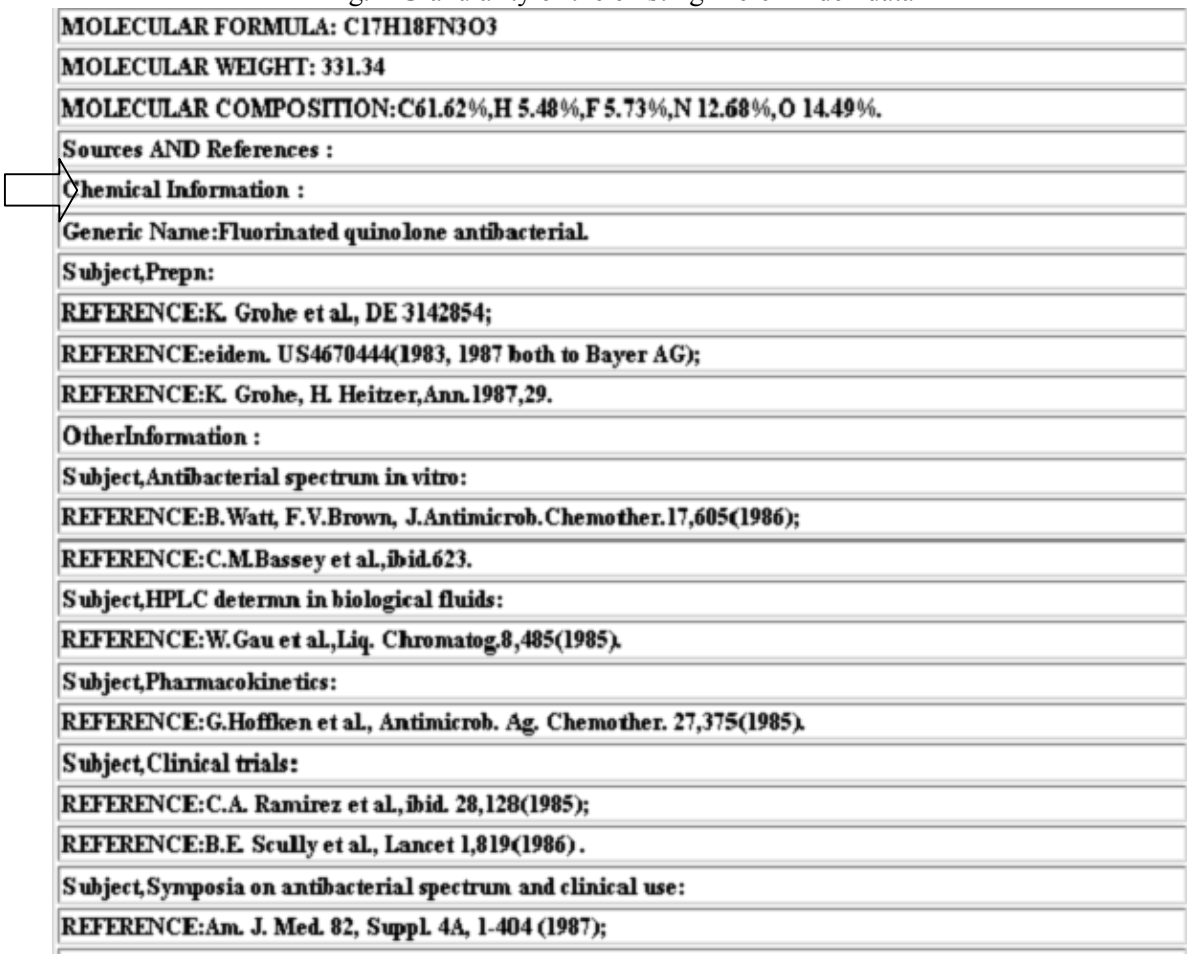
1Retrieving via Web Services
2Retrieving and Picking up target digital objects directly from the global digital library
3Web Services for the global digital library
4Web Services for existing digital library
5Topic Maps for the global digital library
6Topic Maps for existing digital library
7Replying via Web Services

Figure 4 shows “granularity of the existing Merck Index data”, and Figure 5 shows “granularity of the Global Digital Library Based-SDO Hierarchy Structure”. Compare the two parts of “Sources and References”, because Figure 5 is organized by atomic granularity, it can meet high-precision search. Figure 6-8 demonstrate the Global Digital Library services based Web Services, and Figure 9-10 for reifying relation between SDOs by Introducing Topic Maps. Figure 11 is the XML Topic resource of Figrue 9.



Merck Index
THE MERCK INDEX Monograph Number:2377
Chemical Properties
Chemical Properties, THE MERCK INDEX Monograph Number:Ciprofloxacin
Chemical Properties, CAS REGISTRY NUMBER:85721-33-1
Chemical Properties, CHEMICAL NAME:1-Cyclopropyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid;
SYNONYMS(Alternate names and/or trademarks (capitalized) of title compound):
MOLECULAR FORMULA: C17H18FN3O3
MOLECULAR WEIGHT: 331.34
MOLECULAR COMPOSITION:C61.62%,H 5.48%,F 5.73%,N 12.68%,O 14.49%.
Sources/References
LITERATURE REFERENCES:Fluorinated quinolone antibacterial. Prepn: K. Grohe et al., DE 3142854;eidem. US4670444(1983, 1987 both to Bayer AG); K. Grohe, H. Heitzer,Ann.1987,29. Antibacterial spectrum in vitro: B.Watt, F.V.Brown, J.Antimicrob.Chemother.17,605(1986);C.M.Bassey et al.,ibid.623. HPLC determn in biological fluids: W.Gau et al.,Liq. Chromatog.8,485(1985). Pharmacokinetics: G.Hoffken et al., Antimicrob. Ag. Chemother. 27,375 (1985). Clinical trials: C.A. Ramirez et al.,ibid. 28,128(1985); B.E. Scully et al., Lancet 1,819(1986). Symposia on antibacterial spectrum and clinical use: Am. J. Med. 82, Suppl. 4A, 1-404 (1987); J. Antimicrob. Chemother.26, Suppl. F.3-193(1990). Review of clinical safety and efficacy in children: R. Kubin, infection 21,413-421(1993)
PATENTINFORMATION:DE 3142854;US4670444
PATENTINFORMATION, Patent Number:DE 3142854;US4670444
Physical Property Information:Dec 255-257°C
Therapeutic category (In humans):THERAP CAT: Antibacterial

Fig. 4 Granularity of the existing Merck Index data



MOLECULAR FORMULA: C17H18FN3O3
MOLECULAR WEIGHT: 331.34
MOLECULAR COMPOSITION:C61.62%,H 5.48%,F 5.73%,N 12.68%,O 14.49%.
Sources AND References :
Chemical Information :
Generic Name:Fluorinated quinolone antibacterial.
Subject,Prepn:
REFERENCE:K. Grohe et al., DE 3142854;
REFERENCE:eidem. US4670444(1983, 1987 both to Bayer AG);
REFERENCE:K. Grohe, H. Heitzer,Ann.1987,29.
OtherInformation :
Subject,Antibacterial spectrum in vitro:
REFERENCE:B.Watt, F.V.Brown, J.Antimicrob.Chemother.17,605(1986);
REFERENCE:C.M.Bassey et al.,ibid.623.
Subject,HPLC determn in biological fluids:
REFERENCE:W.Gau et al.,Liq. Chromatog.8,485(1985).
Subject,Pharmacokinetics:
REFERENCE:G.Hoffken et al., Antimicrob. Ag. Chemother. 27,375(1985).
Subject,Clinical trials:
REFERENCE:C.A. Ramirez et al.,ibid. 28,128(1985);
REFERENCE:B.E. Scully et al., Lancet 1,819(1986).
Subject,Symposia on antibacterial spectrum and clinical use:
REFERENCE:Am. J. Med. 82, Suppl. 4A, 1-404 (1987);

Fig. 5 Granularity of the Global Digital Library Based-SDO Hierarchy Structure

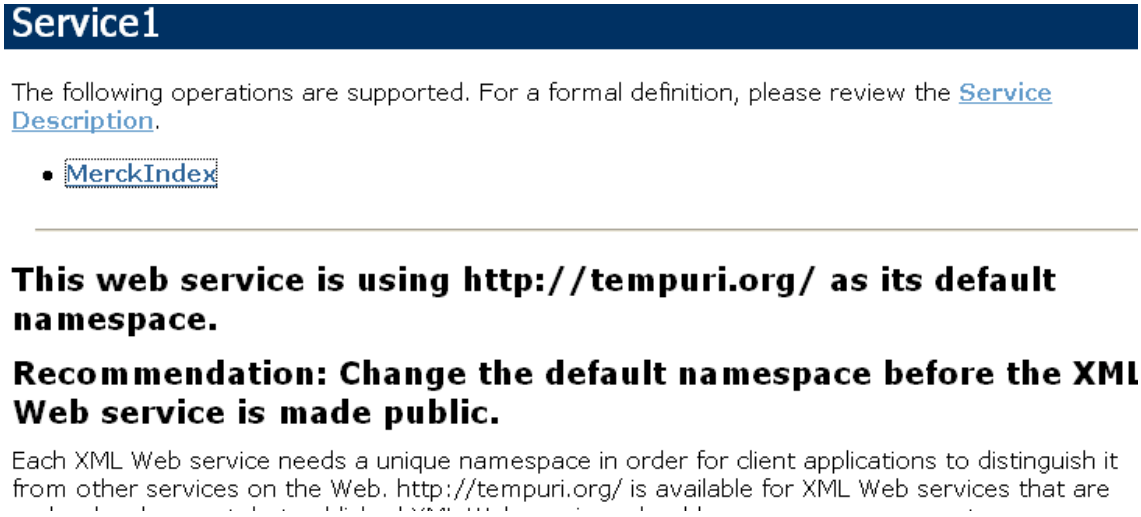


Fig. 6 Web Services for Merck Index

Fig. 7 Invoking Merck Incdex

Fig. 8 Invoked Result

Fig. 9 Describing/Expressing SDOs Relationships by Topic Maps

```
<?xml version="1.0" encoding="utf-8" ?>
- <Patent
    xmlns:xsi="http://www.w3.org/2001/XMLSchema
    -instance">
    <Patentee>K.Grohe et al.</Patentee>
    <PatentNumber>DE 312854</PatentNumber>
  </Patent>
```

Fig. 10 Resources Discovery by Topic Maps

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="TopicMapsExample.xsl"?>
<topicMap xmlns:xlink="http://www.w3.org/1999/xlink">
 <topic>
  <baseName>
  <baseNameString>Ciprofloxacin</baseNameString>
  </baseName>
  <occurrence>
  <instanceOf>
  <topicRef xlink:href="TopicMapsPatentsOccurrence.xml"/>
  </instanceOf>
  <resourceData>Patents</resourceData>
  </occurrence>
  <occurrence>
  <instanceOf>
  <topicRef xlink:href="TopicMapsJournalsOccurrence.xml"/>
  </instanceOf>
  <resourceData>Jurnals</resourceData>
  </occurrence>
</topic>
```

5 Conclusion and Prospect

Metadata means many different things to many different people, and its effectiveness depends on implementation resolving key issues, including: granular (high-precise search), associative, reuse digital object and adaption for Web Services and Topic Maps.

This paper provides a method for naming, identifying, and/or invoking digital objects in a system of the global digital library that provides great flexibility and is will suited to other field , such as E-learning. It allow the possibility of locating digital objects to use Topic Maps . It also admit value-added convertions that various users may use to their

own advantage.

References

1. http://www.ukoln.ac.uk/metadata/
2. J. L. Hunter. A survey of metatata research for organizing the Web. Library Trends/fall 2003, Pages 318-344.
3. http://www.rlg.org/preserv/joint/day.html
4. http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html
5. H. Ishizuka, Y. Li “Digital Library System Based on Concept of Computer-processable Digital Objects ”, Proceedings of Digital Library - IT Opportunities and Challenges in the New Millennium, pages 315-326, 2002.7, Beijing.
6. Y. Li, H. Ishizuka. Reseach and Development of a Web Retrieval System on the Concept of “Computer-Processable Structured Digital Object Based on XML”. Journal of Japan Society of Informatin and Knowledge. Vol.12, No.4, pages 53-68, 2003.