

电子文件管理元数据元素语义的结构化探索^{*}

孙丰满

摘 要 电子文件管理元数据元素语义的结构化是表征元素语义构成部分及其关系的过程,目的是为了实现在元数据元素语义结构的标准化。这种结构化主要包括语义结构化和语法结构化。元数据元素语义的结构化是最基本的语义结构化,是保证元数据基本元素和子元素理解的一致性、实现基本元素和子元素互操作的关键。标准化的元素语义结构有助于实现文件管理元数据描述电子文件的准确性,保证电子文件真实性、可靠性、完整性和可利用性,保证电子文件的法律证据价值。图7。参考文献4。

关键词 电子文件管理 元素语义 语义结构化 元数据

分类号 G254

ABSTRACT In this paper, the author provides a definition of the structurization of element semantics for electronic file management metadata. He thinks that the structurization mainly includes semantic structurization and grammatical structurization, and it is the key for the consistency of the understanding of basic data elements and sub-elements and for the realization of the interoperability of basic elements and sub-elements. Standardized element semantic structures are helpful to the realization of the accuracy of the file management metadata to describe electronic files, can guarantee the truth, reliability, completeness and usability of electronic files, and can guarantee the value of electronic files as legal evidences. 7 figs. 4 refs.

KEY WORDS Electronic file management. Element semantics. Semantic structurization. Metadata.

CLASS NUMBER G254

电子文件管理元数据元素语义的结构化就是表征元数据元素语义构成部分及其关系的过程。所谓表征,其中“表”就是指对语义及其关系进行揭示、表示或表达,“征”就是指对揭示出的语义构成部分及其关系加以标识。

电子文件管理元数据元素语义的结构化的目的就是实现元数据元素语义结构的标准化,具体来说其意义在于:有助于实现在不同空间、时间上的各类用户及计算机对同一元数据元素语义理解的一致性,实现语义内容上的互操作;有助于实现在不同空间、时间上的各类用户及计算机对同一元数据关系语义理解的一致性,实现关系语义上的互操作;有助于实现在不同空间、时间上的各类用户及计算机对同一元数据结构理

解的一致性,实现语义结构上的互操作,从而有效地实现电子文件管理元数据描述电子文件的准确性,保证电子文件的真实性、可靠性、完整性和可利用性,保证电子文件的法律证据价值。

1 电子文件管理元数据元素语义的结构化

1.1 元数据基本元素语义的结构化

元数据基本元素是构成元数据的主体、核心,其语义的理解是对元数据语义理解的关键,因此,元数据基本元素语义的结构化一直是国际电子文件管理元数据领域研究的重点。在元数据发展的早期,元数据标准化程度很

^{*} 本文为国家社科基金重点资助项目“基于XML电子文件管理元数据标准研究”(项目批准编号:04ATQ002)系列研究成果之一。

低,所以元数据基本元素语义的结构化仅是元素最基本的构成部分——定义的揭示。定义的揭示只能实现用户简单的理解,却无法实现计算机对基本元素语义的理解。随着电子文件管理元数据标准化的发展,基本元素的语义构成得到多方面的揭示,既方便了用户的理解,也实现了计算机的理解。从这个意义上讲,元数据基本元素语义的结构化可以分为两种类型,一类是简单的语义结构化,一类是标准的语义结构化。

(1) 简单的基本元素语义结构化

所谓简单的基本元素语义结构化就是简单地揭示元数据基本元素的定义,而对于元数据基本元素其他方面的语义则不加以揭示与标识。如国际电子文件管理领域中最早的元数据研究项目之一——加拿大不列颠哥伦比亚大学的电子文件管理元数据研究项目(UBC项目)^①,其基本元素结构化如图1所示。

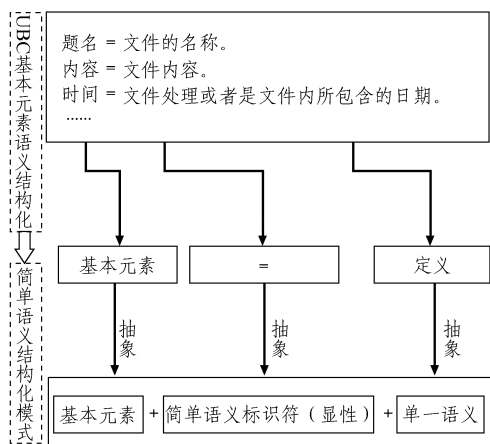


图1 UBC 简单基本元素语义结构化模式示例

从图1看出,元数据的基本元素“题名、内容、时间”等,它们的语义仅仅通过单一语义“定义”来揭示;而语义标识仅仅是通过简单的标识符“=”来表达。由此可见,该结构模式实际上就是:

简单语义结构化 = 基本元素 + 简单语义标识(显性) + 单一语义

简单的基本元素语义结构化的另一个经典例子就是美国匹兹堡大学的元数据研究项目——文件管理中法律证据性功能要求项目(BAC项目)^②,其元素语义结构化如图2所示。

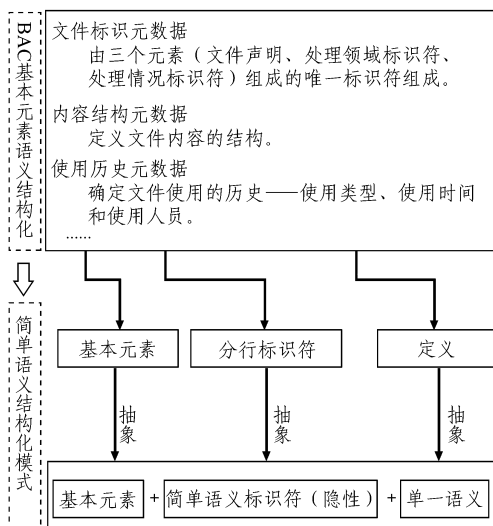


图2 BAC 简单基本元素语义结构化模式示例

从图2看出,匹兹堡项目中基本元素语义的结构化,也是通过对元素语义定义进行揭示并通过换行标识符来实现的,即隐性标识。虽然在“标识”上与UBC项目“=”标识形式不同,但是本质上也是形式的简单语义标识,因此,这种基本元素语义结构化的模式可概括为:

简单语义结构化 = 基本元素 + 简单语义标识(隐性) + 单一语义

由此可见,简单基本元素语义结构化主要是通过对单一语义——定义进行揭示,通过简单语义标识符进行标识,其局限性很显然;①在标识基本元素的语义方面,由于仅仅是采用简单的标识符,并不揭示元素语义,且不规范,因此标准化程度比较低。②在揭示基本元素的语义方面,没能实现多维揭示,只能一维地揭示单

① Luciana Duranti, Terry Eastwood. The Preservation of the Integrity of Electronic Records, 1997.

② The Pittsburgh Project. Functional Requirements for Evidence in Recordkeeping, 1995.

一语义,因而没有充分揭示基本元素的语义。

由于以上两点,所以又造成:①在用户理解方面,由于简单语义结构化对基本元素的语义揭示程度低,难以实现对基本元素的全面理解,造成用户对基本元素语义理解障碍,从而影响元数据基本元素语义在不同空间、不同时间上各类用户理解的一致性,不利于实现不同空间、不同时间上对基本元素语义的互操作。②在计算机理解方面,由于简单的元素语义结构化未能揭示计算机理解所需的语义构成,造成计算机理解的障碍,从而影响了元数据基本元素语义在不同空间、不同时间上各类计算机理解的一致性,不利于实现计算机在不同空间、不同时间上对基本元素语义的互操作。

(2) 标准的基本元素语义结构化

与简单的基本元素语义结构化相对的是标准的语义结构化。所谓标准的基本元素语义结构化,其基本元素语义构成均是按语义标准来进行揭示与标识的。标准的语义结构化不仅揭示了基本元素的定义,还揭示了基本元素的基本属性、标识、关系及管理等诸多语义构成,既实现了用户的理解,也方便了计算机的理解。从目前国际电子文件领域中所制定的电子文件元数据标准来看,基本元素标准的语义结构化主要都是按照国际标准化组织所制定的标准《ISO11179-3:2003,信息技术 元数据元素的规范与标准化 第三部分:元数据的注册与元素的基本属性》来揭示并标识的,如《南澳大利亚电子文件管理元数据标准》,其语义结构化模式如图3所示。

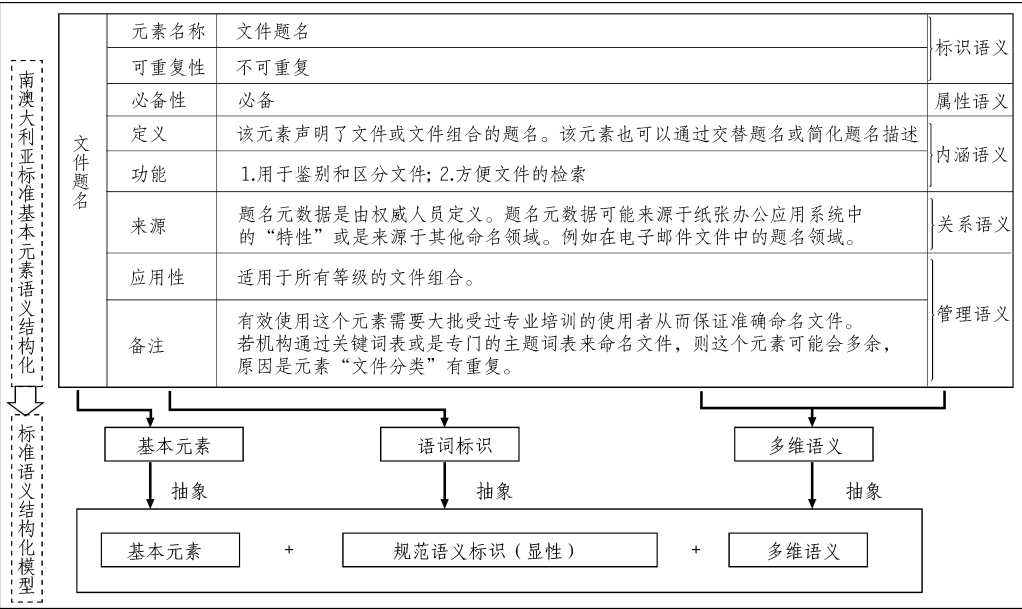


图3 标准的基本元素语义结构化模式示例

以图3中基本元素“文件题名”为例,其语义的构成从元素名称、可重复性、必备性、定义、功能、来源、使用条件、应用性和备注等九个方面来揭示,并由“元素名称、可重复性、必备性、定义、功能、来源、使用条件、应用性和备注”等来加以标识。这样标准的元素语义结构化模式

实际表达了基本元素语义的五个维度,即“标识语义、属性语义、内涵语义、关系语义、管理语义”,从而达到了多维揭示。该基本元素语义结构化的标识采用规范的语义标识——语词标识,不仅揭示了语义的内容,并且十分直观。因此,这种标准语义结构化模型可以概括为:

标准语义结构化 = 基本元素 + 规范语义标识(显性) + 多维语义

与简单的语义结构化相比,标准的基本元素语义结构化,其语义的标准化程度高,具体特点是:①在标识基本元素的语义方面,以国际标准为依据,采用规范的语词标识,揭示了基本元素语义的内容,并使用规范的格式,标准化程度高。②在揭示基本元素的语义方面,能够对基本元素的标识、定义、关系以及管理语义进行揭示,实现了多维语义揭示,结构化程度高。

根据以上优点可以得出:①在用户理解方面,标准的语义结构化由于其结构化程度高,可以全面地理解元素的语义,扫清了用户对基本元素理解的障碍,有利于实现基本元素语义在不同空间、不同时间上各类用户理解的一致性,从而有利于实现不同空间、不同时间上对基本元素语义的互操作。②在计算机理解方面,标准的语义结构化揭示了计算机理解所需的语

义构成,有利于实现元数据基本元素语义在不同空间、不同时间上各类计算机理解的一致性,从而有利于实现计算机在不同空间、不同时间上对基本元素语义的互操作。

1.2 元数据子元素语义的结构化

元数据子元素是基本元素的下位元素,也是元数据元素的重要组成部分。因此,也存在对元数据子元素语义如何结构化的问题。从子元素语义结构化程度的高低来看,可以分为内嵌子元素语义结构化和标准子元素语义结构化。

(1) 内嵌子元素语义结构化

内嵌子元素语义结构化是指对子元素语义的构成内嵌在基本元素的结构中加以表征。内嵌子元素语义结构化有很多种形式,但最典型的是《英国的电子政务元数据标准》,如图4所示。

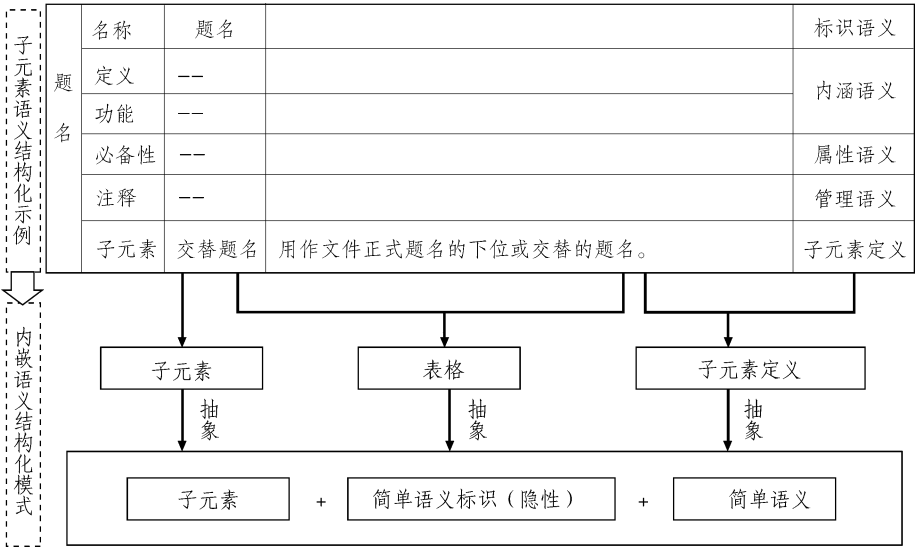


图4 内嵌子元素语义结构化模式

从图4看出,子元素“交替题名”的语义结构是内嵌于基本元素“题名”中的,它仅仅揭示了内涵语义——定义,即是“用作文件正式题名的下位或交替的题名子元素”,但作为揭示内涵定义的语义标识却被隐含起来没有显示,因此

这种内嵌子元素语义结构的模型可以抽象归纳为:

内嵌子元素语义结构化 = 子元素 + 简单语义标识(隐性) + 简单语义

这种子元素语义结构化的模式存在着局限

性:由于内嵌于基本元素结构中,空间上的局限性使其难以对子元素的多维语义作全面揭示,结构化程度低。由此导致:子元素语义揭示简单,无法全面、准确地理解,影响了不同用户和计算机在不同空间、不同时间上对子元素理解的一致性,从而不利于实现不同用户、不同计算机在不同空间、不同时间对同一子元素的互操作。

(2)标准子元素语义结构化

标准子元素语义结构化是将子元素作为一

个独立的元素,通过规范的语义标识实现多维子元素语义的揭示。目前国际文件领域,标准子元素语义结构化也是根据国际标准《ISO11179-3 2003 信息技术 元数据元素的规范与标准化 第三部分:元数据的注册与元素的基本属性》对语义进行揭示和标识的。以《澳大利亚政府电子文件管理元数据标准》中子元素“创建时间”为例,可以看出这类语义的结构化,如图 5 所示。

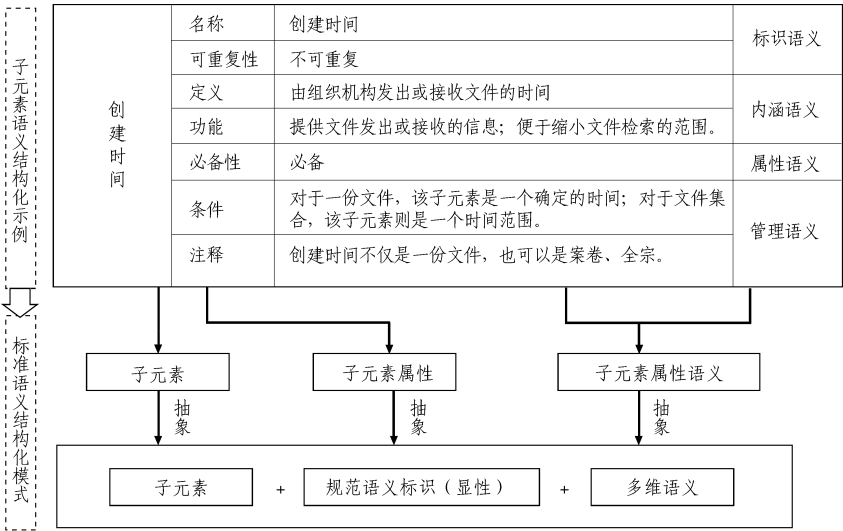


图 5 标准的子元素语义结构化模式

从图 5 可知,子元素“创建时间”是作为一个单独的元素加以语义结构化,并未内嵌于“时间”基本元素结构中。由此看出,标准子元素语义结构化是对子元素通过规范的语义标识——词语标识进行标识,通过多维的语义描述进行语义揭示,其模式可以概括为:

标准子元素语义结构化 = 子元素 + 规范语义标识(显性) + 多维语义

由此可见,标准子元素语义结构化优越于内嵌元素限定词语义结构化:由于子元素是独立于基本元素的,因此有足够空间可对子元素的多维语义作全面揭示,结构化程度高。进而可得:子元素语义揭示充分,可以全面地、准确地对其加以理解,有利于不同用户和计算机在

不同空间、不同时间上对子元素理解的一致性,从而有利于实现不同用户、不同计算机在不同空间、不同时间对同一子元素的互操作。

2 元数据元素语义结构化是元数据元素语法结构化的前提与基础

元数据元素语义结构化程度的高低,影响着用户和计算机对元素语义理解的一致性,决定着不同计算机系统之间对同一元素互操作,从而决定着元数据元素语法结构化程度的高低,因此,元数据元素语义结构化是元数据元素语法结构化的基础。

以图 1 基本元素“题名”为例,其简单语义

结构化仅仅揭示了“题名”的定义,而没有揭示其他方面的语义,因此,在进行语法结构化的过程中,定义基本元素“题名”就会产生语法上的不一致,如以 XML 为例,对“题名”元素进行语法结构化,如图 6 所示。

从图 6 看出,元素“题名”的 XML schema 仅仅定义了元素名称,即 name = “题名”,这样,元素“题名”的语法结构化程度就很低,就有很大的局限性:

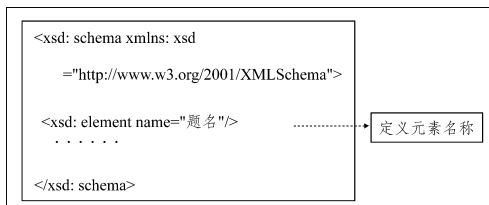


图 6 基于“题名”简单语义结构化的 XML 语法结构化示例

①由于在简单的“题名”元素语义结构化中没有揭示元素的必备性,因此在以元素语义为基础的语法结构化中,就难以依据“题名”元素语义结构进行必备性的统一定义,这样就会导致在不同的计算机系统中对必备性有不同的理解,如计算机 A 系统认为“题名”元素的必备性是“必备”的,那么,“题名”元素就必须出现;而计算机 B 系统认为“题名”元素的必备性是“可选择”的,那么,“题名”元素就可以不出现,这样,同一“题名”元素在不同计算机系统中就会具有不同的必备性,导致不同计算机系统对“题名”元素理解的不一致,从而不利于不同计算机系统之间对相同文件的互操作,难以保证文件的法律证据价值的实现。

②由于在简单的“题名”元素语义结构化中没有揭示元素的可重复性,因此在以元素语义为基础的语法结构化中,就难以依据“题名”元素语义结构进行可重复性的统一定义,这样在不同的计算机系统中对“题名”元素的可重复性就会有自己的理解,如计算机 A 系统认为“题名”元素的可重复性是“不可重复”的,那么,定义“题名”元素出现次数为 1 次;而计算机 B 系统认为“题名”元素的可重复性是“可重复”的,

那么,定义“题名”元素的出现次数为多次,这样同一“题名”元素在不同计算机系统中也会具有不同的重复性,从而导致不同计算机系统对题名元素理解的不一致,也难以保证文件的法律证据价值的实现。

同样,以前述的图 3 “文件题名”为例,其标准语义结构化不仅对元素“定义”进行了揭示,而且还对元素的功能、必备性、可重复性、注释等进行了揭示,因此,在元数据元素语法的结构化过程中,定义基本元素“题名”的语法就可以达到一致,如以 XML 为例,对“题名”元素进行语法结构化,如图 7 所示。

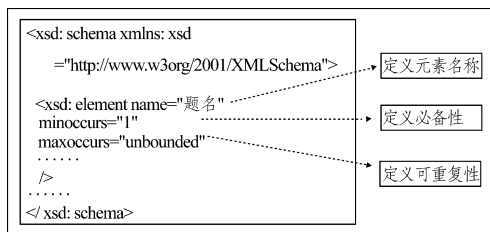


图 7 基于“题名”标准语义结构化的 XML 语法结构化示例

从图 7 看出,元素“题名”的 XML schema 不仅定义了元素名称为“题名”,即 name = “题名”,并且定义了元素的必备性为“必备”,即 minoccurs = “1”,而且也定义了元素的可重复性为“可重复”,即 maxoccurs = “unbounded”,同时,还定义了元素的注释,即 < annotation > 和 < documentation >。这样基于题名元素标准语义结构化的语法结构化,其语法结构化程度就很高,就有很大的优势:

①由于在标准的元素语义结构化中定义了“题名”元素的必备性,因此在以元素语义为基础的语法结构化中,就可以依据“题名”元素语义结构进行必备性的统一定义,这样在不同的计算机系统中就可以实现对“题名”元素在必备性的理解上的一致性,从而有利于不同计算机系统之间对相同文件的互操作,保证文件的法律证据价值的实现。

②由于在标准的元素语义结构化中定义了“题名”元素的可重复性,确定了元素出现的次

2008 信息服务创新国际学术研讨会在中国人民大学召开

2008 年 12 月 21 日, 2008 信息服务创新国际学术研讨会在中国人民大学信息资源管理学院召开。信息服务创新是图书馆学情报学领域的重要问题。近年来, 信息服务实践发展迅猛, 由此带来的很多理论和实践问题, 引起众多国内外专家的关注, 也成为图书馆学情报学教育改革关注的焦点之一。举办此次学术研讨会的目的是加强国际间的学术交流, 跟踪国际前沿研究动态, 把握信息服务创新的国内外理论与实践。中国人民大学信息资源管理学院院长赵国俊致开幕辞, 张甲、曾蕾、秦健、林夏、张东明、王雪茅等 6 位国外专家以及武夷山、周庆山、周晓英、卢小宾等 4 位国内专家在会上做了主题发言, 国防科技信息中心总工程师曾民族致闭幕辞。

此次研讨会吸引了大量来自高校、国家级信息机构和图书馆的人员, 有美国霍普金斯大学、雪城大学、爵硕大学、肯特州立大学的学者、国内 10 所高等院校的信息管理学院的专家以及国家图书馆、中国科技信息研究所、国家科学图书馆、国防科技信息中心、《中国图书馆学报》、中国人民大学图书馆的代表参会。会议主题涉及国家信息基础结构、e-science 和 e-research、开源软件、数字产权、电子交易模式、信息机构服务、术语注册、网络服务系统、机构仓储、企业技术创新、数字图书馆、图书馆情报课程、政府网站信息服务、信息学术交流模式以及图书馆发展等众多内容。(周晓英)

数, 因此在以元素语义为基础的语法结构化中, 就可以依据“题名”元素语义结构进行可重复性的统一定义, 这样在计算机 A 系统和计算机 B 系统对题名元素的可重复性就可以在理解上达成一致, 实现了计算机系统对元素理解的一致性, 从而有利于不同计算机系统之间对相同文件的互操作, 保证文件的法律证据价值的实现。

③由于在标准的元素语义结构化中定义了“题名”元素的注释, 进一步明确了题名元素的应用语义, 因此在以元素语义为基础的语法结构化中, 就可以依据“题名”元素语义结构在语法上进行注释的定义, 这样在人读方面就可以实现对题名元素在理解上达成一致, 实现不同用户理解的一致性, 保证文件的法律证据价值的实现, 同时, 也可以保证元数据在语法应用上的准确性, 从而有利于实现计算机系统对元素理解的一致性, 进而实现不同计算机系统之间对相同文件的互操作, 保证文件的法律证据价值的实现。

参考文献:

- [1] National Archives of Australia. Recordkeeping Metadata Standard for Commonwealth Agencies[S], Version 1.0, May 1999.
- [2] Public Record Office of UK. Requirements for Electronic Records Management Systems—Metadata Standard[S], 2002 revision.
- [3] Government of South Australia. South Australian Recordkeeping Metadata Standard[S], version 2.4.
- [4] International Standard Organization(ISO). Information technology-Metadata registries(MDR) -Part 3: Registry meta-model and basic attributes[S], 2003.

孙丰满 南京政治学院上海分院军事信息管理学系硕士研究生。通讯地址: 上海市黄兴路 2053 弄 211 栋 304。邮编 200433。

(收稿日期: 2007-09-07)