新数字图书馆中的数字服务——表示图书馆目录的新方式

当图书馆终端用户通过联机目录检索某一作者的作品时，他通常会得到一个长长的列表，包括这个作者不同译本和不同版本的所有作品，并按标题或发行日期排序。作为一种尝试，为了对这种混乱状态加以排序，Pode 项目应用了一种基于 MARC 记录所包含信息的自动 FRBR 化的方法。

该项目还尝试用 RDF 表示 MARC 记录，以演示一个作者的全部作品如何能用一个简单和清晰的唯一作品列表呈现出来。结果是作为一种启发，继续将图书馆目录完全转换成 RDF 表示。

Asgeir Rekkavik
Oslo Public Library, Deichmanske bibliotek
Oslo, Norway
E-mail: asgeir.rekkavik[at]kul.oslo.kommune.no

Kim Tallerås
Oslo and Akershus University College of Applied Sciences
Oslo, Norway
E-mail: kim.talleras[at]hioa.no

Anne-Lena Westrum
Oslo Public Library, Deichmanske bibliotek
Oslo, Norway
E-mail: anne-lena.westrum[at]kul.oslo.kommune.no

中文翻译：*刘华梅（中国国家图书馆）*
Chinese Translator*: LIU Huamei (National Library of China)*

Meeting: **80 — *Inspired moments in cataloguing* — Cataloguing**

简介

挪威首都奥斯陆的新图书馆建筑计划，以及关于这些计划对图书馆服务造成的影响的讨论，间接地给奥斯陆公共图书馆提供了机会，以检查他们的元数据怎样在新环境中使用，并能以多种方式提供更好的服务。

在过去的几年中，Pode 项目一直尝试用描述性元数据进行相关的混搭，参考模型如 FRBR（书目记录功能需求的 IFLA 工作组，1998 年），新一代的 OPACs 和关联数据。这一工作至少得出了一个核心思想：一个图书馆仅仅基于现有的元数据，而不是基于提供的元数据的质量，是不能创造更好的服务的。

Pode 项目一年前结束了，但结果和发现已经被进一步用在其他项目中为图书馆用户创造可用的公共服务。该项目基于 RDF 表示 MARC 记录，目的是创建一个交互站，以便以新的方式表示图书馆目录。该项目将链接图书馆目录和外部数字资源，并通过触摸屏界面以可视化和直观的方式呈现出来。我们也在制作一个链接 RDF 编目记录的图书推荐语义集合。

Pode 项目

Pode 项目是用从图书馆目录中发现元数据这种新方式运作的，特别是混合目录内容和其他“混搭”资源。另外，该项目还洞察了当今系统和实践提供的网络技术带来的可能性和挑战性。

该项目使用图书馆集成系统供应商提供的传统协议，同时转换和调整数据实现 FRBR 化以提供“含义层”的关联图书馆数据。该项目还尝试用 RDF 表示 MARC 记录，以演示一个作者的全部作品如何能用一个简单和清晰的唯一作品列表呈现出来，另一方面还能很容易地以不同的内容表达和载体表现浏览。另外，通过链接数据集实例匹配和对应外部集实例，用作者和作品的其它附加信息使表现形式更加丰富。

我们第二个核心思想是：RDF 表示我们图书馆目录给了我们更多可操作的数据集，比传统 MARC 数据集有更多机会为我们的用户创造服务。

FRBR 和 FRBR 化——通过图书馆点击列表发现

克努特·汉姆生(Knut Hamsun,1859-1952)是挪威最杰出的小说家和三位挪威诺贝尔文学奖得主之一。他的文学作品包括小说 30 余部，几个剧本和短篇小说集，一部诗歌集和一些非小说和传记作品。汉姆生的所有产出一共是 40 部作品，这是图书馆用户很容易浏览的一个书目。然而，要满足图书馆用户的想象是完全不同的。在奥斯陆公共图书馆的联机目录中，限定检索作者“Hamsun, Knut”，将产生 568 条命中结果（2012 年 5 月 7 日获取）。这对于一个写了 40 部作品的作者所提供的命中结果当然是太多了。注意这还是高级限定检索的结果。而大多数图书馆用户尝试的将是一个更通用的简单检索，会提供更长的结果列表。

联机目录的问题是不能区分同一作者的不同作品和同一作品的不同版本。在命中的 568 条记录中，有多达 63 条是对应一部小说“hunger”的不同表示方式，有不同版本，不同格式和不同语言的译本。用户必须能选出他们想要的图书，有声图书或电影，还必须能选出他们想要阅读的图书是什么语言，但是对大多数用户来说，当 OPAC 让他们在 20 多种挪威语的不同版本的“Hunger”中做出选择时，会觉得更困扰而不是更有用。

图书馆标准

图书馆编目历来注重描述物理实体。一本书的每种载体表现用一条单独记录表示，并且描述同一作品载体表现的记录间没有功能性连接。一个图书馆用户，用一特定题名检索联机目录，或许可以在候补书单上做标记，从而借到一部经典小说的特定版本，而不考虑大量的已经存在的同一本书的其他版本。另一位用户可能最终没有得到所有想要的书，却无意中捡到一个不再有可用副本的版本。Pode 项目实验基于这样一个假设，图书馆用户通常感兴趣是找出某个特定题名，而不是该题名的某一特定版本，这种兴趣尤其对于小说作品，因为不同版本通常含有完全相同的内容。从图书馆系统的角度来看，用户理想的是预订一个题名，而不必选择不同的版本。当然，那些在意版本的人也应该有机会仔细选择，但是为什么要强迫所有人去挑选呢？

正如很多人所指出的，目前的图书馆标准是在网络和当前生成、分配和利用元数据的基础设施之前制订的。另外，将图书馆数据引进到电子领域的标准也是在实体关系模型(ER)和关系数据库发明前几年制订的。

这就为实现像 FRBR（将版本和作品分离）这种参考模型带来了挑战，它是基于 ER 分析和 MARC 记录中不能实现的关系。

MARC 格式体现了技术里程碑和卡片目录的逻辑性，这被用来开发自动化。卡片目录的元数据由人工读取和解释，这一特点在 MARC 格式中延续，这就决定了制作单条记录主要由人类可读的文本字符串组成。当然，这只是图书馆元数据实践的简单描述。制作 MARC 记录过程在像 AACR2 和 ISBD 这样的编目规则中详细描述了其复杂相互作用，文本字符串隐含的大多数意义在这些规则里都能找到。在网络环境中，我们希望机器来处理数据并解释给我们。文本字符串必须绝对一致以便机器能准确地解释数据，并创建一个有用的数据表示形式。

在关系数据库和关联数据环境下，最佳实践学说是提供唯一标识符——分别用主/外键和 URIs，以避免歧义。MARC 记录缺乏这样明确的标识符，能帮助我们的索引工具和搜索引擎区分具有相同名称的两个作者，或者合并那些具有相同的唯一标识作品可能表示的是同一人的作者。当我们希望机器能识别文献、作者和不同 FRBR 实体之间的关系时，这些类型的挑战当然会继续。哪些书代表了一个特定内容表达或一部作品的载体表现？我们如何识别像短篇小说这样的作品，当他们只在 MARC 说明字段以文本描述的时候？对于大多数 1970 年之后出版的图书，我们有识别载体表现的 ISBN 号，我们有能从计算机追溯到的姓名和题名，但是没有一致的目录和机器可读标识，这仍是一项艰巨的任务。它至少需要一个适当的清理过程。

MARC 记录的 FRBR 化

Pode 项目使用由挪威科技大学开发的一个工具，将 MARC 记录自动实现 FRBR 化（Aalberg, 2006）。这一工具在 MarcXchange 格式的输出编目中使用 XSL 转换来对书目记录中的数据进行分类，这正是 FRBR 实体数据加以应用的基础。

图书馆编目记录主要是载体表现的描述；单条记录描述某一出版作品的特定版本，如出版时间和地点，物理描述，ISBN 等具体载体表现信息。 但是记录也包含用于内容表达和作品载体表现的相关信息。例如，数据包含的作者名字和一本书的原始题名是描述作品实体的一些信息，而有关文献语言和格式的信息也是对于内容表达的一些说法。FRBR 化工具将确定哪些领域适用于哪些 FRBR 实体，并用它们之间的 FRBR 关系将每条记录分成作品一部分，内容表达一部分，以及载体表现一部分。如果该工具为两个不同的 MARC 记录输出相同的作品描述，就假定这些记录描述的是同一作品的两个不同体现方式。该工具也会产生 2 个和 3 个实体，如机构和主题。

最初，该项目通过自动 FRBR 化系统在选择的没有清理过的记录语料库运行（908 条记录，描述 Hamsun 和挪威另一个作者，Per Petterson 的载体表现）。第一次尝试的结果不尽如人意。这主要是由于 MARC 记录的缺失信息和编目实践不一致。为了进一步改进实验，项目不得不清理一定数量的记录，以便得到充分表达和一致的数据。

清理过程

简单的说，清理过程主要包括识别和为缺失记录增加原始题名和统一题名，为收集在一卷里的单一作品和短篇小说增加信息，设置标识符以区分有意义作品题名和无意义题名。另外，还有一些时间花在纠正错别字和错误上。总共花了大约 60 小时清理 Knut Hamsun 和 Per Petterson 的作品，包括用于确定校正规则的时间。

清理过程中的主要工作之一是识别和改进记录，翻译那些完全缺少挪威原始题名的作品，或是只在人工阅读的说明中提供了原始题名的作品。另一项工作是，无译名的翻译题名和原始题名不同的情况下，要增加统一题名。由于挪威的拼写革命，Hamsun 的一些作品已经在不同的时间以不同的题名发表。例如，短篇小说 *Paa tourney*(原始题名)也以 *På tourné* 和 *På turné* 的形式发表过。

另一项任务是设置标识符以区分真正的有意义作品题名和无意义题名。 无意义的题名通常是在一个作者的作品合集中出现，把内容放到一起并由别人而不是作者本人给出版物定名。这类出版物不应该被列为 FRBR 化的书目，即使集合中的个别小说或短篇小说应该被列出[1]。

例如：

a) 245 10 ‡aGrowth of the soil ‡cKnut Hamsun ; W. Worster 从挪威语翻译

（Hamsun 原著的英语翻译。第一个指示符设为 1 表明这是一个有意义的作品题名）

b) 245 00 ‡aTales of love and loss ‡cKnut Hamsun ; Robert Ferguson 翻译

（英语短篇小说集合，原本不是一起出版的。第一个指示符设为 0 表明这是一个无意义题名）

关于清理过程的详细信息，请参阅附录。

[1] 虽然 MARC 字段之间的关系是通过记录隐式表达的，不容易混淆，并且能明确表达由几部作品组成的文献的各字段之间的关系。

结果

虽然第一次尝试实现 Hamsun 的 149 部作品 FRBR 化，但对 MARC 记录清理后进一步减少到了 84 条。而 Per Petterson 的情况，我们看到作品数量从 14 增加到 41，原因是为一些编目记录增加了单个短篇小说和散文的题名。作品的结果列表几乎完全对应两位作者的实际书目。唯一的例外是 Hamsun 的一部作品是由挪威国家图书馆 Hamsun 参考书目列出的，而在奥斯陆公共图书馆的馆藏中是缺失的。因此，清理过程无意中为我们识别馆藏缺失作品提供了方法。否则你处理数百个载体表现的列表，这是一个繁琐的过程。

RDF 化

FRBR 化数据集转换成 RDF，用 XSLT[2]和项目开发的 MARC 字段和 RDF 属性词之间的一个对照表。对照表主要使用众所周知的词表和本体，如 DC 元数据语词，书目本体，核心 FRBR，FOAF 和 SKOS。但该项目还构建了一些更具体的子属性，比遵循这些词表能更准确地表达我们的数据。后期该项目发现 RDA 词表[3]包含很多属性词，覆盖了很多需要被表达出来的图书馆专业信息。在之后对对照表的修订中一些已经取代了我们自己的属性词。

完整的对照表请见 http://www.bibpode.no/blogg/?p=1573。

数据一旦被转换成 RDF，通过和其他数据集链接将变得更加丰富。作品可以和 DBpedia 和 Gutenberg 项目的实例链接，人物可以和 DBpedia 和 VIAF 链接。为了能够按作品年代排序，该项目新增了作品实例的第一版日期信息。而 MARC 记录即使全部都包含，这一信息也很难从中提取出来。有了这个新的和丰富的数据集，该项目就能开发一个网络应用程序，允许终端用户通过浏览图书馆这些作者作品的全部馆藏，从中选择一个简短的作品列表，而不是搜索平台所列的数以百计的载体表现。另外，应用程序还可以为终端用户提供作者来自 DBpedia 的相关信息，还可以链接到 Gutenberg 项目的数字全文版本。

开发了一个简单的网络应用程序，允许终端用户浏览图书馆的部分馆藏，以 FRBR 实体聚合，通过关联数据[4]从可用的外部资源提供额外的信息。

用 RDF 表示两位作者很快给了我们更多可用的数据集。经验是积极的，我们决定将整个图书馆目录转换，看看能否实现我们最基本的愿望：为我们的用户创造更好的服务。

关联数据和 RDF 作为新服务的基础

链接奥斯陆公共图书馆的“图书推荐”和“活动的书架”项目，已投入大量的工作把图书馆目录转换成 RDF 格式。此时，我们投入这一领域的努力是更加详细和全面，比我们在 Pode 项目做

[2] 以后的这种工作，我们已经用 Ruby 脚本 YAML 映射文件代替了 XSLT，详见 https://github.com/bensinober
[3] http://rdvocab.info/
[4] http://bibpode.no/linkedauthors/

的实验能提供更多的可能性。奥斯陆公共图书馆目录包括大约 42 万条记录，可追溯到 20 世纪 70 年代末。

RDF 是一种使图书馆编目数据真正实现机器可读的格式。MARC 格式的机器可读性意味着计算机能读取、存储和处理编目记录中的字符，RDF 的机器可读性意味着计算机能读取我们数据的实际含义或语义。简单的说，图书馆数据转换成 RDF 是把编目记录“转换”成简单的陈述集，在这里我们可以用唯一标识符表示我们想说的一些事情，也可以表示我们想要给这些事情的一些信息。

我们能思考新想法的同时不忘记我们的老技能吗？我们不是世界上第一个把图书馆数据转换成 RDF 的图书馆。但是当这一领域的大量工作聚集在提取和转换编目记录基本信息的一小部分的同时，我们却试图付出我们的“所有”。图书馆编目员是全面的和无价的，在图书馆工作的任何人都知道，我们可以获取比题名、作者、主题和 ISBN 更多的关于文献的信息，编目是至关重要的。我们试图尽可能地保持 MARC 记录中的信息，并尽可能准确和正确地表达出来。我们通过 NORMARC 标准逐字段的系统的工作，同时对于怎样解释数据，怎样把内容以 RDF 表达出来都需要我们做决定。

通过 YAML 映射把二进制 MARC 转换成 RDF 的完整转换脚本见 https://github.com/digibib/marc2rdf。

同时，重要的一点是编目的 RDF 版本不是简单地变成另一种只是语言不同而其它完全相同的数据集。编目规则和 MARC 格式是几十年前设计的，在一定的限制下可能能达到最好的效果，有些限制在这些标准制订时是相关的而对今天来说是不相关的。此外，许多规则的特点是由描述文献和定位文献这两个双重功能决定的。如果旧规则和旧思想带进新格式中，那么旧的限制也将陪伴它们。RDF 格式给了我们 MARC 缺乏的可能性，我们利用这个优势是最重要的。

一个特定例子：当一条编目记录陈述卡尔 • 马克思是“共产党宣言”的作者，而恩格斯是合著者时，是什么意思？这种功能性区别对于两个人的真实功能没有任何实际意义。实际上在 MARC 所有字段中并没有一个作者字段；只有“主要款目”字段。这一字段的功能是说文献在集合中的定位。实体文献只能有一个位置，因此它只能有一个主要款目。马克思的名字在扉页上列在恩格斯名字的前面；因此，他单独获此殊荣，作为这本书的作者被说明，而恩格斯的部分被减少为“合著者”。

当我们用 RDF 表达这一信息时，我们不必关心这种人为的和非功能性的区别。卡尔 • 马克思是这本书的作者，恩格斯也是。这些陈述之间没有冲突。

另一个例子是当编目记录信息模糊不清和依赖语境时。以这条附加款目为例：

*700 0 $a King, Stephen
$d 1947-
$e forf.
$j am.
$t Rita Hayworth and Shawshank Redemption

这个例子中编目文献和中篇小说"Rita Hayworth and Shawshank Redemption"之间有什么联系呢？这完全依赖语境，如果编目文献是一本书，这种陈述很可能意味着这篇中篇小说是文献的一部分。如果编目文献是一个 DVD 光盘，这很可能意味着电影基于中篇小说中的故事。数据的正确解释就需要知道文本和电影有关的东西以及它们是怎样关联的。换句话说，需要一个人类读者。

这两个例子告诉我们一些有关 MARC 格式所能表达的局限性。第一个例子说明，诸如主要款目和附加款目这样的概念并不是设计来描述图书、电影或音乐的主要东西。它们的目的是组织实体馆藏和选择一种方式允许图书馆用户查看卡片目录。第二个例子影响的是机器的可读能力。当数据的正确解释依赖人类常识时，数据也很难被机器真正读取。

SPARQL 是我们用来访问 RDF 数据的查询语言。我们可以用 SPARQL 语言做任何普通的目录检索，还能得到很多额外的信息。

CCL 图书馆检索为我们提供了许多方式组合查询，以获得高级的和专指的检索。但是有一个你永远无法逃避的限制：无论你如何构建你的检索，返回的结果都是编目记录。你可以问任何你喜欢的问题，但只能以“哪本书……”开头。

你不能要求特定语言的文献主题列表，或者特定主题的作者列表。要回答这样的问题，你必须找到图书馆有那种语言或那一主题的那些图书，然后进行繁琐的过程，通过列表以识别不同主题或作者。

传统图书馆检索另一个限制是他们不能包含未知的实体。例如你不能构建一个 CCL 检索，使它能返回一个给定题名的作者的所有书籍。你必须先查找已知题名，找出作者的名字，然后再检索该作者的书籍。这或许并不是个苛刻的要求，但是如果你需要的信息依赖于多个未知实体呢？像“我能在图书馆找到哪些小说，也能找到改编本的 DVD？” 这种提问，对传统图书馆检索工具是不可能回答的，除非你有足够的时间。有了 SPARQL 的帮助，你可以这样提问：

给我这样的文献，文献-1 和文献-2，满足：

文献-1 是一个 dvd 电影；

文献-2 是包含小说的图书；

文献-1 基于作品 X；

文献-2 是作品 X 的载体表现。

结束语

我们将继续用 RDF 表示我们目录的工作。下一步将实现检索，看看结果是否比我们此时从 OPAC 上得到的要好。你可以在 http://digital.deichman.no/访问我们的进展。

附录——清理和校正

主要工作是确保所有的转换记录（由于外国语言和挪威语改革）在 240 字段包含原始作品题名。当 240 缺失时，基于 574 注释字段的信息自动添加，574 字段是 NORMARC 包含原始题名信息的特定字段。当说明字段也缺失或自动转换失败时，需要人工补充题名。

另一个耗时的工作是在 245 字段和 740/700‡t 字段识别和确定有意义作品的真正的（原始的）题名，并根据修正规则设置恰当的指示符。

处理 740 字段的第二个指示符，只有当字段包含一个有意义的作品题名时，该项目选择用 2 标识。对于所有其它题名指示符设为 0，即使他们是分析款目。即使这种做法不符合用 700‡a + ‡t 表示分析题名款目的约定，但这一决定是清楚地检测分析作品出版的最有效方法。

基于 FRBR 化的首次尝试结果，修正的 Hamsun 和 Petterson 的 908 条记录包括：

- 修正 5 条记录的 008 字段的语言编码；
- 为 85 条记录的 240 字段添加统一题名（严格按照 NORMARC 术语是“原始题名”），修正了 24 条记录的 240 字段存在的错别字；
- 修正了 6 条记录 245‡a 的错别字(或错误 ISBD 语法)；
- 修正了 245 字段的第一个指示符：之前修正 137 条记录的指示符 1 为 0 或空，同时指示符 1 为 1 的用在 774 字段，之后修正 263-651 的分布；
- 修正 700 字段：在原始记录中，我们发现 948 条 700‡a 和 545 条 700‡t 字段。修正后的记录数量分别减少为 917 条 700‡a 和 481 条 700‡t。这种变化是由于对同一作者的所有记录更系统的使用了 740 字段（这在 100 注册登记）。
- 改变 700 字段的第二个指示符以明确一个款目是否是唯一作品。

参考文献

Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. *ERCIM News*, (66). Retrieved from http://www.ercim.eu/publication/Ercim_News/enw66/aalberg.html