FRBR 化：利用 UNIMARC 连接字段识别作品

Manolis Peponakis
Michalis Sfakakis
Sarantos Kapidakis
数字图书馆和电子出版实验室
爱奥尼亚大学档案和图书馆学系
克尔基拉(科孚岛)，希腊

中文翻译 / Chinese Translators
金文昌、朱青青 / JIN Wenchang & ZHU Qingqing
中国国家图书馆 / National Library of China

Meeting: **187 — *Advancing UNIMARC: alignment and innovation* — IFLA UNIMARC Programme (UNIMARC)**

摘要

此研究的主要目标是结合 MARC21 FRBR 化的实践与 UNIMARC 格式语义，并突出 FRBR 化背景下 UNIMARC 和 MARC21 两者间的一些差异。主要的重点是确认利用 UNIMARC 连接字段识别书目记录功能需求(FRBR)作品实体的可能性。在我们的方法中，我们提出所有用 45X 字段连接的记录可能属于同一个其记录也包含这些字段的作品。我们使用来自希腊高等学校图书馆编目部的古希腊作家记录的一个样本，作为这个方法的一个测试集。

FRBR

FRBR 是一个由 IFLA 研制的概念上的实体关系模型。实体间基本关系的图示法如下图 1 所示。

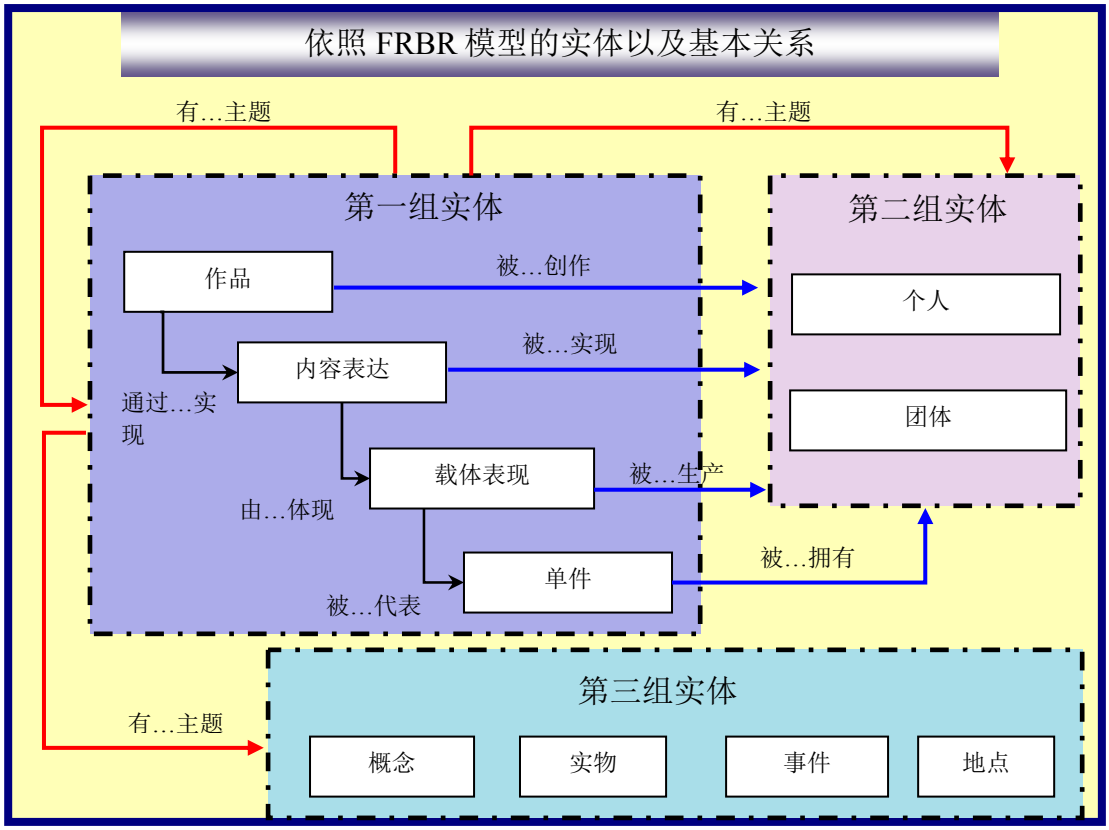


图 1：实体以及基本关系(基于 Manguinhas 等人的图示法，2010)

FRBR 是一个实体关系模型，是书目领域的一个普遍看法，意图独立于任何编目规则。FRBR 既不是一个元数据方案，也不是编目规则。AACR 的继承者资源描述与检索(RDA)是一个贯彻 FRBR 的编目规则。

由于这篇论文的重点是第一组实体，图 2 尝试给出一个有关第一组实体含义的简例。

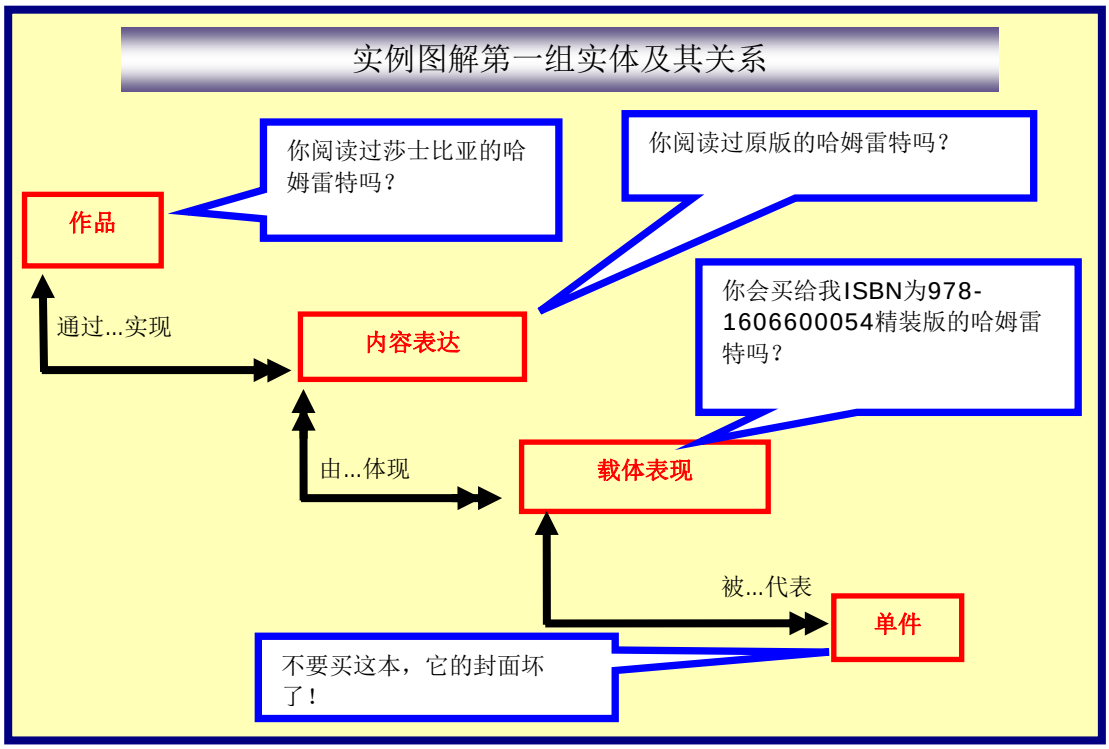


图 2：实例图解第一组实体及其关系(Peponakis 等人，2010)

FRBR 化

传统的目录已达到其极限是普遍公认的，并且正如 Yee 提出的(2005: p.77)：“为了改善系统设计以及 FRBR 化 OPAC 显示和索引，进行更为合理的利用我们现有的几百万条 MARC21 书目记录、规范记录和馆藏记录是必要的。”因此，图书馆应该开发可在不同种类馆藏和元数据方案之间可实施的工具(Naun, 2010: p.333)。FRBR 提供一种对书目数据的当下理解，但正如 Rajapatirana(2005)所述，“重新编目不是我们的选择”。因此对图书馆来说，主要的挑战是：为了提供增值服务，对现有的书目记录的利用。此决定导致发明了允许以新格式重建现有数据的方法。

FRBR 化是利用以前的编目(在其他编码模式中编码)的记录查找和综合 FRBR 实体的过程。为了描述这个过程，Babeu(Babeu, 2008: p. 17)准确地报告了 *FRBR 目录*、*FRBR 的系统*和 *FRBR 执行*这些术语可交替使用，但这些术语没有一个清晰的含义。Babeu 她自己更喜欢“FRBR 启发的目录”这个术语去描述 Perseus 项目背景下的 FRBR 化过程。

任何 FRBR 化尝试的出发点是识别代表一部作品的书目记录，然后是识别这组内的潜在的内容表达和载体表现。作品的识别是最关键的一步，因为它涉及整个数据库并且定义所有的后续步骤。为了顺利进行聚集，利用单独的书目记录并且进行比较，产生若干 “关键码”。相同的关键码意味相同的作品[1]。依照相关书目(Aalberg 2006, Freire 等人 2007, LC FRBR 显示工具)和 FRBR 对作品的定义，一个关键码应该合并三个基本

[1] 该处措词是有争议的，考虑到事实上不同的关键码并不一定意味着不同的作品。关键码之间的差异可通过多种相似的方法测量，在上述基础上设定限制，两条记录将被视为属于同一个作品。

的信息，即作品的著者、作品的题名和资料的类型(例如电影或文本)，通过它们作品能够被表达。

FRBR 第一组实体的产生是基于著者-题名关键码。可应用两种方法产生关键码。在第一种情况下，组成关键码的数据直接取自书目记录。在第二种情况下，利用一个规范文档作为中介。此过程的一个图示法如下图 3 所示，图 3 中虚线指明一个规范文档作为产生关键码的中介[2]。

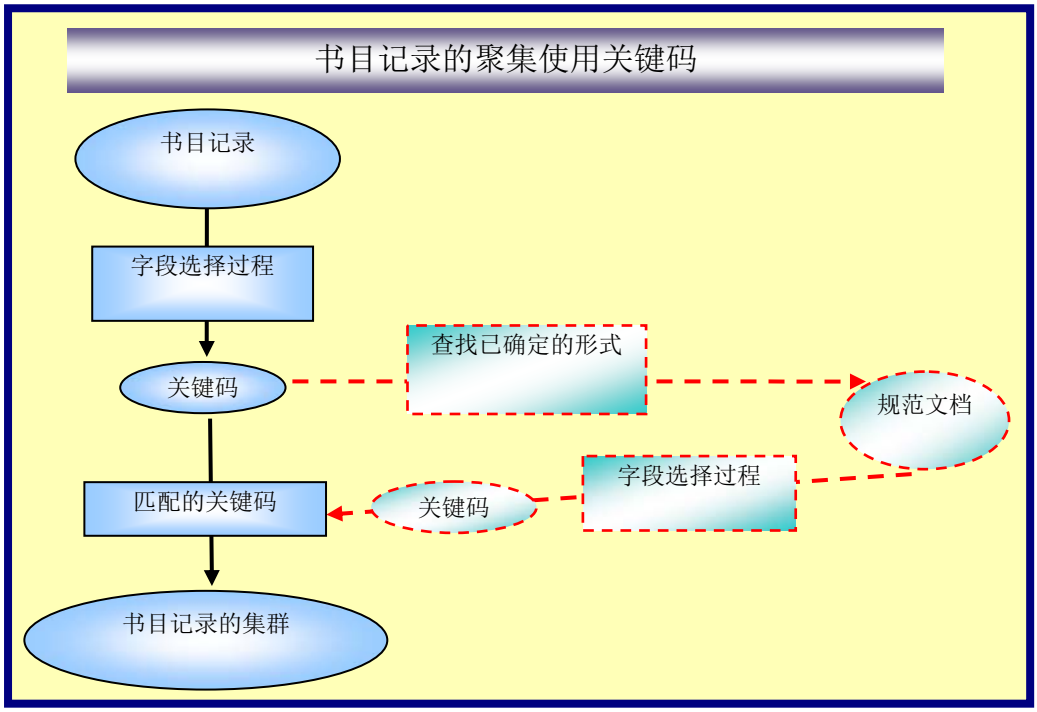


图 3：书目记录的聚集使用关键码的过程(利用或未利用一个规范文档) (Peponakis 等人，2010)

第二种方法的好处是显然的，因为它提供从规范文档获取额外信息的可能性。匹配相同实体的不同的语言表现是有可能的，就“Aristophanes”来说，也以“Aristofanis” 和“Aristophanis”出现。

关键码部分：UNIMARC 和 MARC 21

对于关键码的构成，两个元素在所有情况下是公有的，即题名和著者。利用记录类型能达到进一步的规范。OCLC 的算法不包括记录类型这样的信息，其创建“FRBR 作品集”而不是作品，而国会图书馆(LC)的 FRBR 显示工具则考虑记录类型。在我们的方法中，记录类型也被纳入考虑。因此，我们也利用三个部分构成关键码。第一部分是著者，第二个部分是题名，第三个部分是记录类型。就这三部分来说，UNIMARC 和 MARC 21 两者之间存有语义上的差异。

MARC 21 和 UNIMARC 两者之间关键的差异与是否存在主要款目有关。在 MARC21 的背景下，主要款目是必备的，相反对于 UNIMARC 来说，它是可选的。

[2] OCLC 的算法包括规范文档，但 LC 工具(http://www.loc.gov/marc/marc-functional-analysis/tool.html)不包括。

著者关键码部分

如果一个主要款目著者字段(700、710、720 字段)存在，我们选择这个字段。主要款目不存在的情况下，我们按下列顺序选择一个字段：(Sfakakis and Kapidakis, 2009)

- 第一著者个人名字段，即 701 字段，没有$4 子字段或 $4 子字段有等于“070”(即著者的责任方式代码)的值；
- 第一著者团体或会议字段，即 711 字段，没有$4 子字段或 $4 子字段有等于“070”的值；
- 第一著者家族名字段，即 721 字段，没有$4 子字段或 $4 子字段有等于“070”的值。

正在测试一种启发式的改进，是另外一个规则，即首先研究责任说明 (MARC21 中 245 字段的$f)，然后从以上提及的字段中选择匹配的已确定的名称形式。

题名关键码部分

OCLC 的算法(Hickey and O’Neill, 2005)，在题名字段定义下列选择顺序：

- 统一题名(主要款目) (MARC21 130 => UNIMARC 500, 指示符 2 值 1
- 统一题名(非主要款目) (MARC21 240 => UNIMARC 500, 指示 2 值 0
- 编目员补充的翻译题名 (MARC21 242 => UNIMARC 541)
- 正题名(UNIMARC 200 => MARC21 245)
- 其他变异题名 (MARC21 246 => UNIMARC 517)
- 先前题名(MARC21 247 => UNIMARC 520)

根据它们的定义，UNIMARC 的 45X 连接字段涉及的记录被认为是同一作品的不同的内容表达或载体表现，比如其他版本、译本、复制品。之前的列表不包括连接字段。因此，被连接记录的识别和检索在这个过程中是一个重要的问题。

记录头标类型关键码部分

正如已提及的，关于定义“记录类型”的记录头标，MARC21 和 UNIMARC 两者之间存有语义上的差异。依照 UNIMARC 电子资源指南，有一个利用电子资源(而不是一幅印刷地图)的记录头标值编目数字化资料(以一幅地图为例) 的选项。在此基础上，在如下举例说明的几组中，我们使用值“l=电子资源”。另一方面，MARC21 中，明确了“电子资源的种类是以其最重要的方面编码的 (例如文字资料、图形、测绘制图资料、声音、音乐、移动图像)”。为了聚集同一作品下的具有不同记录类型的记录，我们建议如下的分组。正如下图 4 所示，针对记录类型具有不同值的记录，他们可能属于同一作品，也可能属于不同的作品(见图 5 中具有记录 4、6、8 的例子)。

图 4：基于记录头标的分组提议

一个实例

依照 FRBR，下列三条记录属于同一作品，这部作品由两个内容表达和三个载体表现组成。基于上述，聚集属于同一作品的如下所有记录的关键码，将是“著者=**HOMER** – 题名 = **ILIAD** – 记录类型 = **TEXT**”。

记录 1 – **书**	
题名 / 著者	The Iliad / Homer ; translated by E.V. Rieu
出版	Harmondsworth : Penguin Books , 1954
载体形态	xxv, 466 p., 20 cm.
统一题名	**Iliad**
著者	**Homer**
译者	Rieu, Emile Victor, 1887-1972
正文语种	English
记录 2 – **书**	
题名 / 著者	The Iliad / Translated by E. V. Rieu
出版	Baltimore : Penguin Books , [1964, c1950]
载体形态	469 p., 18 cm.
统一题名	**Iliad**
著者	**Homer**
译者	Rieu, Emile Victor, 1887-1972
正文语种	English
记录 3 – **书**	
题名 / 著者	Ομήρου Ιλιάδα / μετάφραση Ν. Καζαντζάκη, Ι. Θ.Κακριδή
出版	Αθήνα : Εστία, [1997]
载体形态	401 σ., 22 εκ.

统一题名	**Iliad**
著者	**Homer**
译者	Καζαντζάκης, Νίκος ; Κακριδής, Ιωάννης Θ.
正文语种	Modern Greek

表格 1：三条记录构成一部作品、两个内容表达和三个载体表现[3]

建立于连接字段

考虑到 UNIMARC 允许被连接记录的控制号可存在也可不存在这一事实，我们作分别处理。大体上，控制号的存在(与否)与执行的连接技术有关。通常，在嵌入字段技术的情况下，记录控制号存在；而在标准技术的情况下，它不存在。

连接字段嵌入 001 字段的 UNIMARC 记录

在一条被连接记录的控制号存在的情况下，如果记录头标承认它，不管关键码执行的结果如何，所有用 45X 字段连接的记录被认为是属于同一作品。在记录头标分组不同的情况下，他们构成不同的但仍然相关的作品。例如，在图 5 中，记录 4 连接记录 6 和记录 8；但仅记录 4 和记录 6 属于同一部作品，记录 8 不属于是因为它的记录类型不同(关于记录类型分组见图 4)。

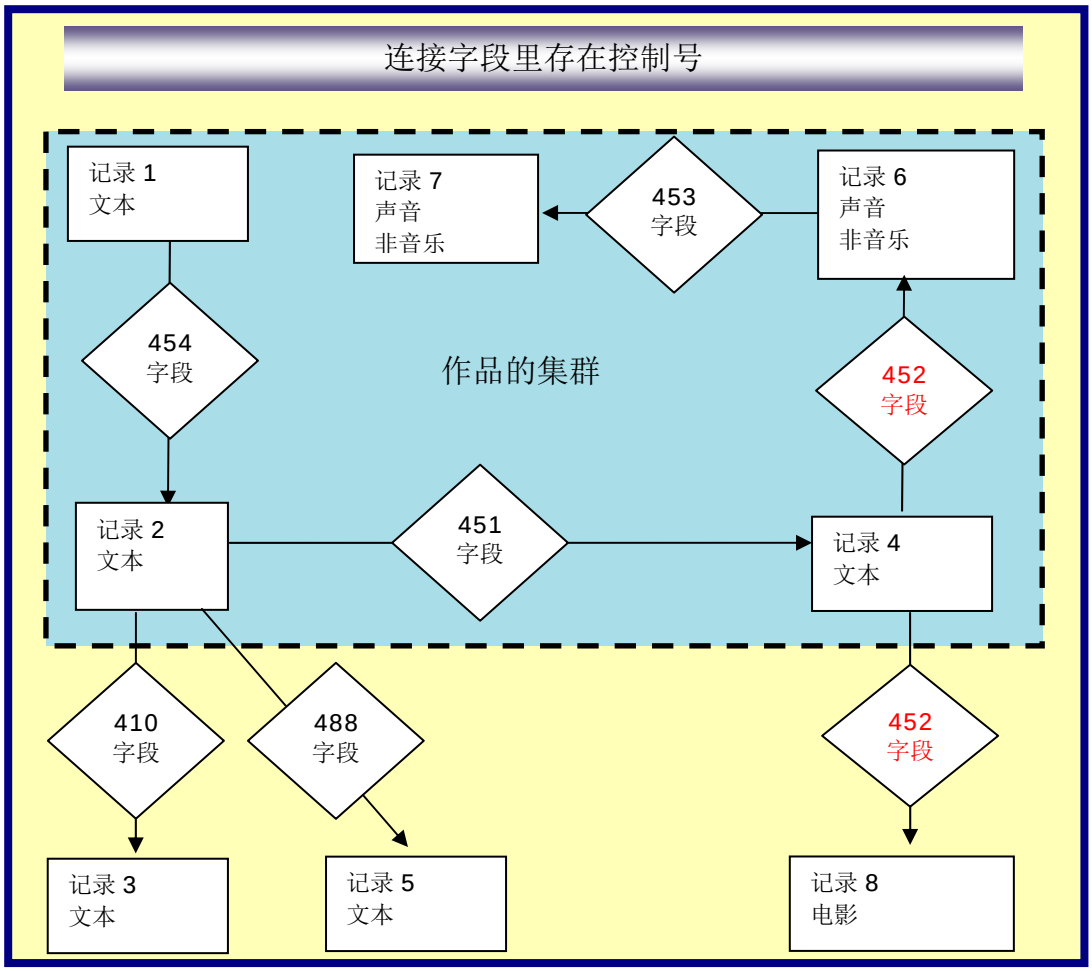


图 5：嵌入 001 字段。淡蓝色背景色部分(包括虚线之内)举例说明该作品

[3] 该作品是荷马的文本的《伊利亚特》，第一个内容表达是 Rieu 翻译的英文版(记录 1 和 2)，第二个内容表达 是 Kazantzakis 和 Kakridis 翻译的现代希腊语版本 。每条记录代表一个不同的载体表现。

连接字段未嵌入 001 字段的 UNIMARC 记录

在这种情况下，连接字段的数据能用来产生关键码。我们观察到，采用标准技术的 45X 字段的信息比 200$a 的信息更正式。实际上，200$a 没有一个具体载体表现的描述，只是一个更为正式的题名(接近统一题名)。如此，使用 451、452、455、456 这些字段而不是 200$a 是更有效的。

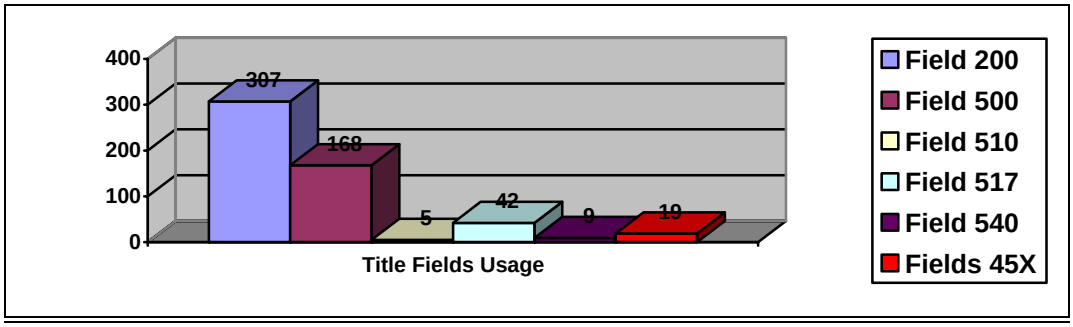
为了定义连接字段的选择顺序(尤其在“453 译为”和“454 译自”的情况下)，我们考虑 101 语种字段。如果指示符是“1=单册是原著或中间作品的译著”，字段“454 译自”被设置在统一题名之下。如果指示符 1 是“0=单册是原著”，我们不使用 453 字段。

评估连接字段用于产生关键码的作用

为了加强连接字段能被用作提高查全效率这一假设，我们设立了一个实验。我们使用来自希腊高等学校图书馆编目部的样本记录作为测试集。这是一个来自 54 个图书馆拥有超过 350 万条记录的大的 UNIMARC 数据库。这个数据库的主要特征是拥有多语种数据、没有共同的规范文档，并且执行的是各自不同的编目政策。

我们选择古希腊作家的作品是因为经典作家的作品拥有许多的内容表达和载体表现，并且能够组成一个用于测试 FRBR 化算法成效的理想“区域”。为了避免有争议的结果，我们从样本中人工排除了所有描述作品片段或多作品装订在一册的记录。

由于希腊高等学校图书馆编目部的连接字段政策在连接字段没有使用 001 字段这样的事实，我们只能采用连接字段被用于构成关键码的方法。首先，我们利用对 OCLC 算法的轻微的修改[4]，去检验以我们的数据集 FRBR 化过程的成效。主要的问题是低的查全率。该算法仅将少数记录聚集在一起。就准确度来考虑，它似乎很有效。



图表 1：题名字段分布

样本由属于 12 部作品的 307 条记录组成。因此，完全的成功是产生 12 个关键码。用于作品识别的最重要的题名是统一题名字段。不幸的是，正如图表 1 所示，大约仅有一半的记录有这个字段。具体来讲，307 条记录全部有 200 字段(100%)；168 条记录有 500 字段(54.7%)；5 条记录有 510 字段(1.6%)；42 条记录有 517 字段(13.6%)；9 条记录有 540 字段(2.9%)和 19 条记录有 45X 字段(6.1%)。仅 3 条记录同时有统一题名和 45X 字段(0.97%)。

为了评估我们算法的成效，我们将单链接聚类应用于产自我们样本记录的两个作品关键码集。第一个作品关键码集由没有使用连接字段产生的关键码组成(基于 OCLC)，而

[4] 我们没有使用一个规范文档，并且我们的元数据是 UNIMARC 格式而不是 MARC21 格式。

第二个集正如上述部分所描述的，使用连接字段。对这两个关键码集进行聚类，85 个聚类产自第一个关键码，78 个聚类产生第二个关键码。

连接字段的使用将作品汇集的成效提高了大约 9%。即使产生的聚类数字的比较，没有单独提供一个对于总体过程的成效的准确指示，在聚类的内容没有被核查和任何聚类仅包含相似记录的情况下，我们看到，两种方法之间，额外聚类的比例是 0.9。此外，从聚类的评估措施例如修正的 RAND 指数和平均侧影宽度信息来看，改进也得到证实。RAND 指数评估正确结果的百分比(正确关键码的匹配)，同时修正的 RAND 增加评估的灵敏性。侧影宽度评估一个关键码有多成功地在聚类，即安置在正确的聚类中。更具体地说，修正的 RAND 指数和平均侧影宽度信息值分别等于 0.56 和 0.81。RAND 指数接近我们的估计，而许多单一聚类的存在影响侧影宽度信息的高度改进。

结论和进一步的工作

首先要弄清楚一点，有时我们碰到的是作品而不是作品集。他们类似 OCLC 的作品集，但就我们的情况，这些作品具有显著的差异：就记录类型来言，他们被更加明确地区分。另外，正如 FRBR 中所述，“是什么构成了一部作品以及一部作品和另一部作品之间的分界线在哪里，事实上不同文化之间的看法可能大相径庭。因此不同文化或国家团体确立的书目规范可能因其用于界定作品的标准不同而有区别” FRBR (p. 16)。

即使有外加连接字段用于关键码的产生，结果显示一个低的查全率。执行不佳的主要原因是缺乏统一题名(500 字段) 以及存在多种多样的正题名。 307 条记录中，有 141 个唯一的正题名(200 字段)，而正如图表 1 所示，总共有 550 个题名字段。每条记录仅利用一个字段，似乎忽视了 243 个题名的重要性，243 个题名构成的数量几乎等于实际使用的数据。为了查全率的显著增加，我们也计划使用这个数据，也就是先前被忽视的题名字段。为了识别作品，我们会将所有的字段彼此比较，而不是仅选择一个题名字段。

参考文献

Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. In**:** *ECDL 2006, Alicante, Spain, 17-22 September 2006*. Gonzalo, J. et al. (eds.) Berlin, Heidelberg: Springer, pp. 453–456. Available at http://www.springerlink.com/content/5356711834963732/fulltext.pdf. [Last accessed 29/05/2011].

Babeu, A. (2008). Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. [Internet] Perseus Digital Library. Available at http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf. [Last accessed 29/05/2011].

Freire, N., Borbinha, J. and Calado, P. (2007). Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques. In: *ICADL 2007, Hanoi, Vietnam, 10-13 December 2007*. Berlin, Heidelberg: Springer, pp. 267–276. Available at http://www.springerlink.com/content/d06r28v440n1x420/.

Hickey, T.B. and O'Neill, E.T. (2005). FRBRizing OCLC's WorldCat. Cataloging & Classification Quarterly. 39 (3/4), pp. 239-251.

IFLA (1998). Functional Requirements for Bibliographic Records. Available at http://www.ifla.org/VII/s13/frbr/frbr.pdf. [Last accessed 29/05/2011].

LC FRBR Display Tool (The Library of Congress' Network Development and MARC Standards Office) http://www.loc.gov/marc/marc-functional-analysis/tool.html.

Manguinhas, H., N. Freire, and J. Borbinha. "FRBRization of MARC records in multiple catalogs." In *Proceedings of the ACM International Conference on Digital Libraries*, 225-234, 2010

Naun, C.C. (2010) "Next generation OPACs: A cataloging viewpoint." *Cataloging and Classification Quarterly* 48 (4), pp. 330-342.

Peponakis, M.; Sfakakis, M.; Kapidakis, S. (2010) "FRBRization: Seeking for the "key" to Works' Identification" (written in Greek). In *Proceedings of the 19th Hellenic Conference of Academic Libraries*. Available at http://library.panteion.gr/19libconf/conference_en.php [Last accessed 29/05/2011]

Rajapatirana, B. and Missingham, R. (2005). The Australian National Bibliographic Database and the Functional Requirements for the Bibliographic Database (FRBR). *The Australian Library Journal*. 54 (1), pp. 31-42. Available at http://www.alia.org.au/publishing/alj/54.1/full.text/rajapatirana.missingham.html. [Last accessed 29/05/2011]

Sfakakis, M. and Kapidakis, S. (2009). Eliminating query failures in a work-centric library meta-search environment. *Library Hi Tech*. 27 (2), pp. 286-307

Tillett, B. (2004). What is FRBR? A conceptual model for the bibliographic universe. [Internet]. Available at http://alia.org.au/publishing/alj/54.1/full.text/tillett.html . [Last accessed 29/05/2011].

Yee, M.M. (2005). FRBRization: a Method for Turning Online Public Finding Lists into Online Public Catalogs. *Information Technology and Libraries*. 24 (3), pp. 77-95. Post print available at http://repositories.cdlib.org/postprints/715/. [Last accessed 29/05/2011]