arXiv:2301.01193v1 [cs.DL] 3 Jan 2023

衡量数字图书馆中数据和元数据的多样性

Rafael C. Carrasco[1*]、Gustavo Candela[1] 和 Manuel Marco-Such[1]

[1*]阿利坎特大学语言与信息系统系，Carretera San Vicente del Raspeig s/n，圣维森特-德尔-拉斯佩奇，03690，阿利坎特，西班牙。

通讯作者。电子邮箱：carrasco@ua.es；贡献作者：gcandela@ua.es；marco@dlsi.ua.es；

摘要

多样性指数传统上用于衡量生态系统的生物多样性，方法是计算物种或物种群的有效数量。与丰度不同，丰度与可用数据量相关，而多样性指数能够更稳健地反映个体的变异性。这类指数可用于数字图书馆，以识别主题分布的趋势、比较不同作者使用的词汇，或分析语义元数据的覆盖范围。

关键词： 元数据、数字图书馆、开放数据、集合即数据

1 引言

丰富度通常定义为生态系统中存在的物种数量，但它对生物多样性的呈现有限，因为它对所有群组一视同仁，不考虑各群组的相对丰度。相比之下，多样性指数 [5]是数值估计量，通过赋予丰度较高的物种更大的权重，同时衡量丰富度和均匀度。因此，它们能够估算物种的有效数量；由于罕见且可能未被检测到的个体所占比重较小，这一估算比样本量更稳健。

随着数字图书馆日益便于访问，人们越来越有必要探讨哪些文献计量指标能够让其特征更易于理解。有人认为 [8]，多样性指数能够有效厘清丰富度与数据量之间的相关性。因此，本文旨在分析多样性指数如何帮助研究人员和专业人士评估数字馆藏内容的词汇多样性以及元数据覆盖范围。

关于文本内容，传统上一直使用形符类符比(TTR)来衡量文档的词汇多样性。TTR 的计算方法是用文本中不同词语（类型）的数量除以词语（词元）的数量。例如，先前的研究比较了不同的方法，包括 MLTD [9] 和$vocd$ [10]，，以评估 TTR 及其在样本中的变异性。一些研究人员 [7] 还探讨了不同体裁是否可以通过特定的 TTR 概率分布来描述。

以往研究建议应用多样性指数来评估文档的词汇丰富度[6]。不过，数字图书馆的其他特征也可以通过多样性概念进行分析。例如，馆藏中主题或作者覆盖范围的局部和

时间变化可以通过计算多样性指数来更好地考察，因为这类指数对那些无法代表馆藏的低频项目不那么敏感。

让我们回顾一下，在生态学中，对于具有N 个群体或物种的生态系统，其真实多样性（或 k 阶多样性指数）定义为

$$D^{[k]} = \left(\sum_{n=1}^{N} p_n^k \right)^{\frac{1}{1-k}} \tag{1}$$

其中， p_n 是第 n类的概率或相对丰度，而参数 k 决定了频繁出现的群体与不频繁出现的群体各自所占的相对权重： k 越大，稀有物种的重要性就越低。

因此，这是一系列指数 $D^{[k]}$，其中最受欢迎的包括香农指数（$k = 1$）和辛普森指数（$k = 2$）[12]。尽管参数 k 会影响所得多样性的值，但在比较不同地点或时间段的多样性时，具体选择哪个参数并不重要。尤其是在研究数字图书馆数据和元数据时， k =1是自然的选择，因为 $D^{[1]}$ 它可以轻松地与信息论中定义的源熵 [13], 联系起来，定义如下

$$H = -\sum_{n=1}^{N} p_n \log p_n$$

因此不难证明，当 k趋近于 1 时，可得 $D^{[1]} = \exp(H)$。我们还注意到， $k = 0$ 会得出样本的丰富度 R 。

本文将探讨多样性指数在分析数字图书馆生成的数据（第 2 节）和元数据（第 3 节）中的适用性。我们将基于图书馆发布的关联开放数据馆藏 [1] 进行图书馆间的比较，因为这些数据提供了一个开放基准。

2 词汇多样性

其词汇中的条目数 M ，也称为词元类型数，可反映文档的词汇多样性。然而，词元类型数取决于文档长度，并且 M 呈现出

随着已处理词元数量单调增长 $n \leq N$ ， N 即文档长度（见图1）。这种无界增长与一个众所周知的事实相符：馆藏中的词元大致遵循齐夫分布 [11]。然而，这使得人们无法直接依据所用词汇的规模比较文本。

图中词元类型的数量可以用一个仅含两个参数的幂函数Cn^{α} 准确近似：尺度参数 C 和指数 α。最适合这些示例的参数列于表 1，并用于绘制图 1中的曲线，这些曲线与数据点紧密吻合。

	C	α
《乌略亚侯爵府邸》	6.7	0.68
Do¯ña Perfecta	6.9	0.66
《伽拉苔亚》	11.1	0.59

表1 图中各曲线的最优参数 Cn^{α} 见图 1。

多样性指数的一个潜在优势是，它们由单个有限值构成，且含义直观。如果已知词汇的基础概率分布（并且该分布是平稳的），就能准确计算类型的多样性；但在实际中，必须根据观察到的频率，从文本样本中估计这些概率。随着文本长度增加，估计的准确性也会提高；随着词元数量增长，结果将趋近真实值。然而，在最常见的情况下，样本量并不足以近似渐近值：如图2所示，文档结束时，香农多样性指数通常仍在增长。

图2中的多样性图要求使用饱和函数来拟合观察到的形态。传统上用于根据不同样本量的样本估计生物多样性的一种函数 [4]是饱和指数函数

$$\Delta_{M1}(n) = D\,(1 - e^{\alpha n}), \tag{2}$$

其中仅涉及两个参数，即指数α 和多样性指数的渐近值 D 。

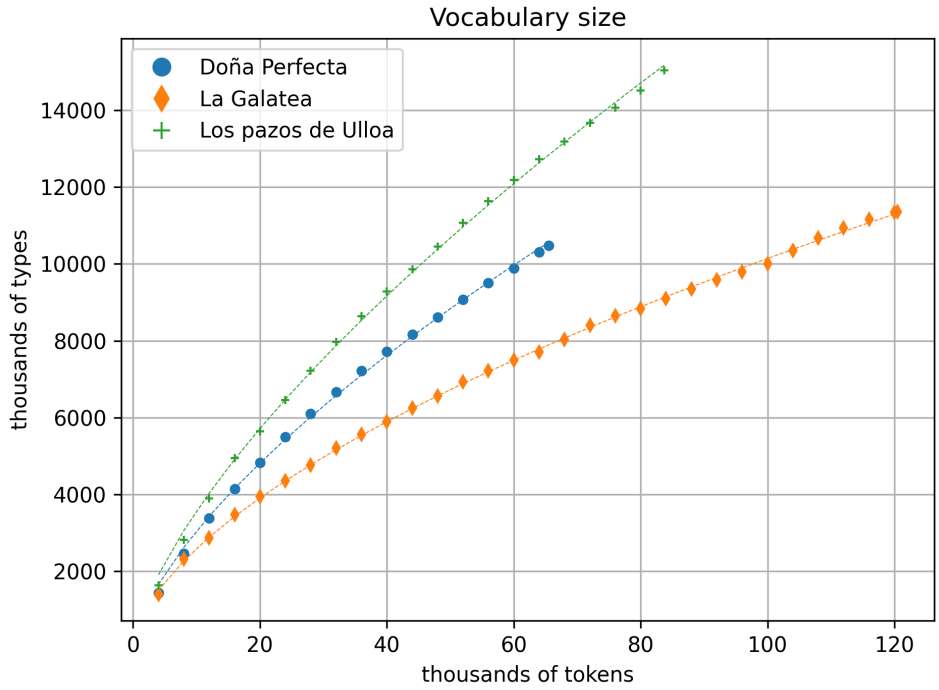

图 1 三部小说的词汇量随阅读词元数量变化的情况： *Los pazos de Ulloa* 埃米莉亚·帕尔多·巴赞的作品、 *Doña Perfecta* 贝尼托·佩雷斯·加尔多斯的作品，以及 *La Galatea* 米格尔·德·塞万提斯·萨维德拉的作品。

第二种传统的渐近模型 [4] 用于物种累积曲线的是二参数函数

$$\Delta_{M2}(n) = D\,\frac{n}{n+c}. \tag{3}$$

在我们的实验中，将模型 M1 和 M2 外推时，它们通常会低估较大样本的多样性。因此，我们研究了其他饱和函数，尤其是广义单项式商

$$\Delta_{M3}(n) = D\left(\frac{n+b}{n+c}\right), \tag{4}$$

以及幂商

$$\Delta_{M4}(n) = D\left(\frac{n}{n+c}\right)^{\alpha}. \tag{5}$$

我们注意到，在所有模型中， D 是渐近值，也就是真实多样性指数。

使用一万个词元对曲线进行外推，以估算更大取值范围内的结果，结果表明模型 M4 始终优于其他模型（见图3）。鉴于预测具有较高的准确度，可以认为，模型 M4 计算出的外推多样性（参数 D 的值）可用于比较文本的词汇多样性，或按作者、体裁或历史时期标注的馆藏的词汇多样性。

我们的结果表明，模型 $M4$ 预测的数值并不取决于样本文本的大小。作为示例，图4 展示了一位多产作者（洛佩·德·维加）的作品词汇多样性随文本长度变化的情况。我们发现的变异性可能与作品风格有关（例如，押韵作品往往表现出更高的多样性），但多样性与作品长度没有显著相关性（皮尔逊 $R \simeq -0.08$）。

3 元数据多样性

3.1 编目记录

多样性指数也可用于分析数字图书馆创建的编目元数据。例如，图5 展示了三家发布了完整编目数据馆藏的图书馆目录中书籍作者的丰富度与多样性

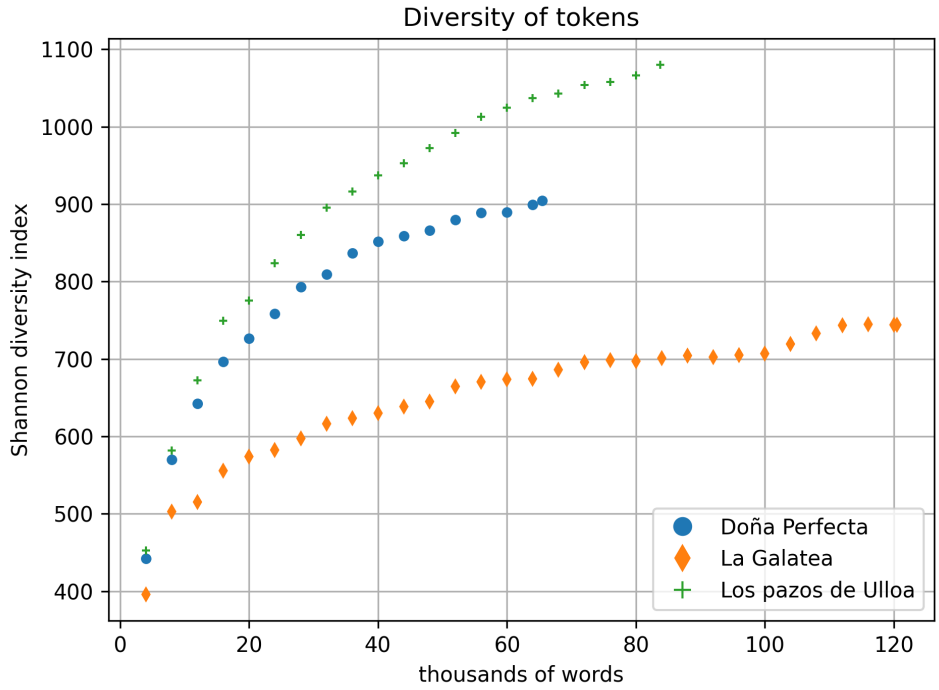

图 2 图中所示作品的香农多样性指数1。

并采用开放许可证：一家大型图书馆（美国国会图书馆，LoC[1]）、一家中型图书馆（根特大学图书馆，UGent[2]）和一家小型图书馆（米格尔·德·塞万提斯虚拟图书馆，瓦伦西亚图书馆[3]）。

丰富度和多样性曲线显示，随着时间推移持续单调增长，没有迹象表明近期会达到平台期。与 LoC 和根特大学馆藏（52–54%）相比，瓦伦西亚图书馆的多样性与丰富度之比更低（约 33%），这反映出其范围较窄——瓦伦西亚图书馆专注于西班牙语文学和历史——表明在其目录中，贡献大量作品的作者所占比例较低。事实上，瓦伦西亚图书馆馆藏中每位作者的平均作品数为 $\mu = 4.9$，而美国国会图书馆（$\mu = 2.5$）和根特大学图书馆（$\mu = 2.1$）的这一平均值更低。

我们还研究了数字图书馆中主题的覆盖范围是否保持稳定，以服务于特定受众；还是倾向于覆盖更广泛的领域。图6 展示了分析完整主题标目字段描述词，以及将其内容拆分为主题、年代、地理或其他细分项时的趋势（例如，描述词“商业–历史”会拆分为两个主题：“商业”和“历史”）。

在所分析的样本中，主题的多样性通常会随着时间推移而持续增长，无论是丰富度还是多样性都是如此。然而，瓦伦西亚图书馆的情况并非如此：将主题细分后，情况便有所不同。一方面，这是因为年代细分的使用更加频繁。另一方面，对记录的检查显示，图书馆在初期之后，逐步提高了历史和文学（尤其是戏剧）领域中西班牙语内容所占的比例，如今这些内容已占其馆藏近三分之一。因此，瓦伦西亚图书馆近来逐渐发展成为一家专业化程度更高的图书馆。

3.2 关联开放数据

在过去十年里，文化遗产机构开始转向采用语义

[1]美国国会图书馆完整图书记录：www.loc.gov/item/2020445551 [2]根特大学图书记录：lib.ugent.be/info/exports [3]米格尔·德·塞万提斯图书记录：data.cervantesvirtual.com/datasets

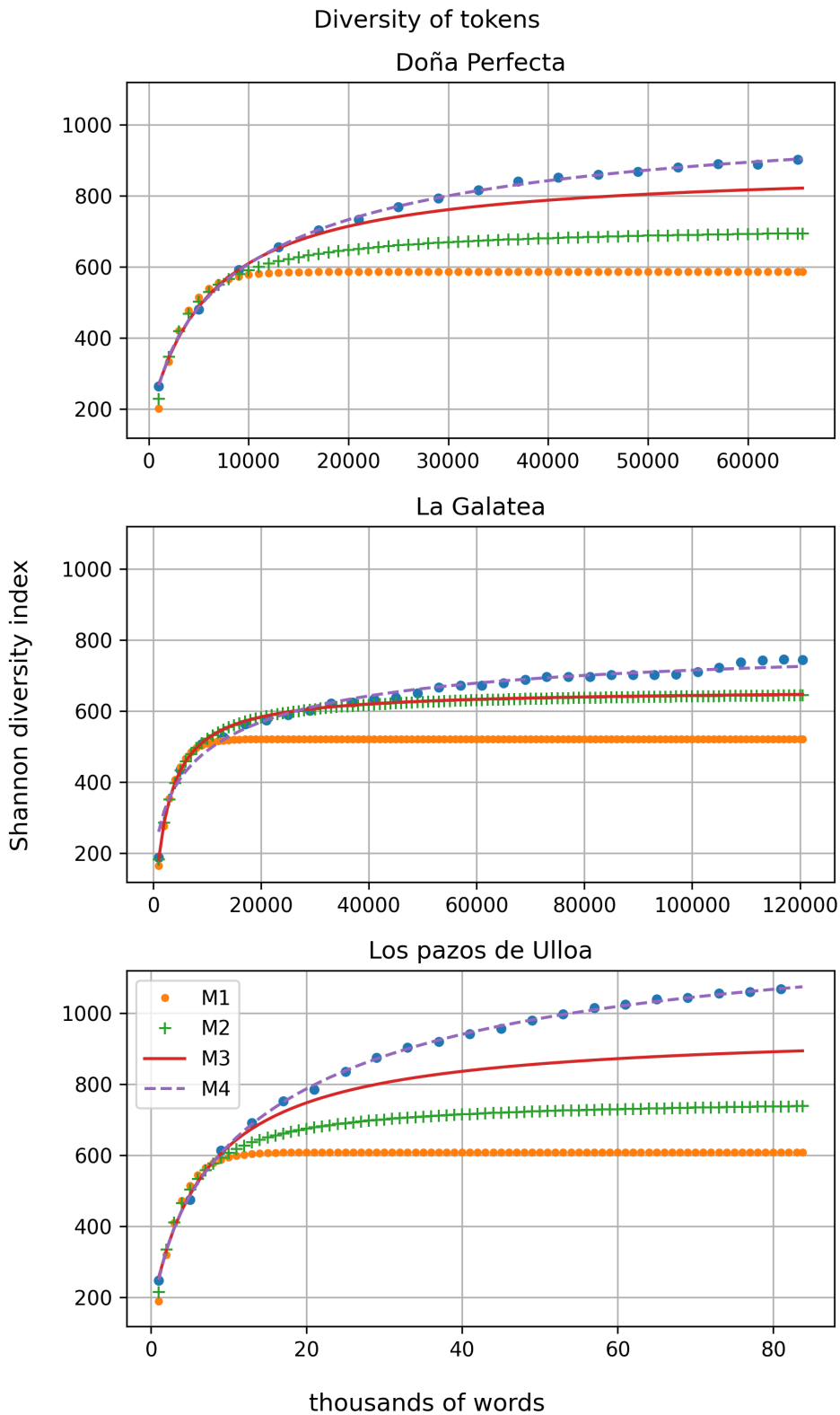

图3 预测　　使用最初的 10000 个词元来确定最优参数时，模型的预测能力　　s.

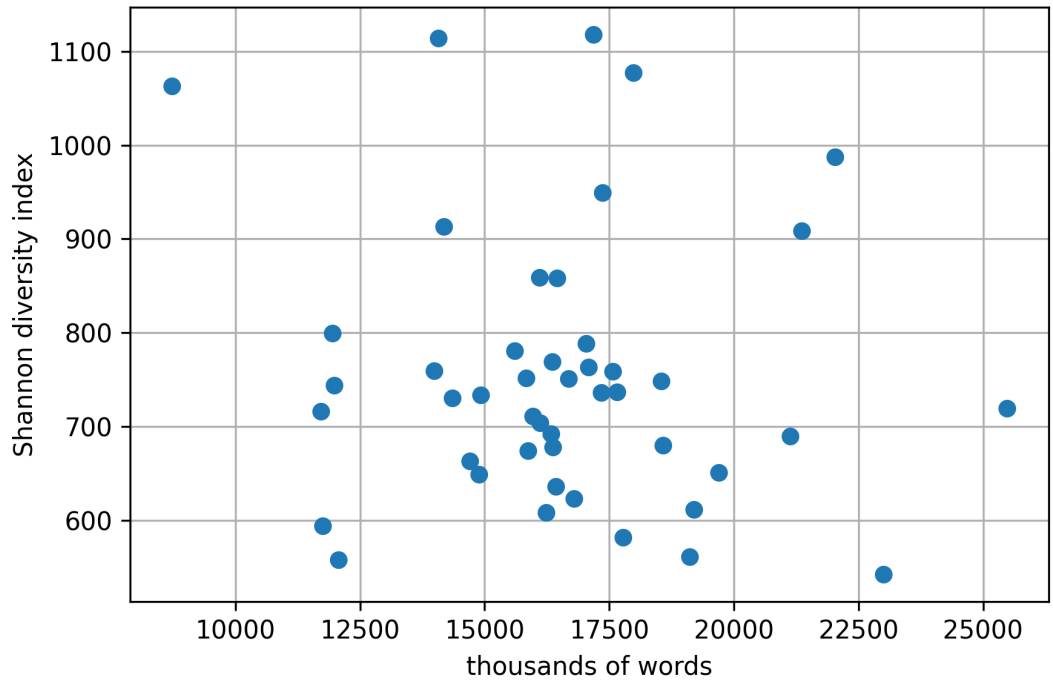

图 4 </style'>洛佩·德·维加作品的香农多样性指数。

网络 [2] 和关联开放数据概念，使用万维网联盟的资源描述框架来表达语义关系 [16]，并使用 SPARQL [15]语言对其进行查询。RDF 通过将资源（图书馆的内容）归类到不同的类中（如人物、作品或名称）来描述资源，并使用属性（如作者）来表达资源之间的关系。资源和属性都通过统一资源标识符（URIs）标识：例如，三元组（X，P，Y）可以将人物的标识符 X 与名称 Y 的标识符关联起来，两者由属性 P连接，其中 URI P 的含义是“名称为”。类似地，形式为（X，rdf：type Z）的三元组声明 X 属于类 Z。

图书馆逐步调整了其目录 [14] ，以便发布关联开放数据（LOD）存储库。不过，如表2所示，它们使用了多种词汇表来定义 RDF 类和属性。这些存储库也以多种形式提供，包括公共 SPARQL 端点、开放档案计划元数据采集协议（OAI-PMH）接口，甚至开放获取转储文件。[4]

为了测试多样性指数在关联开放数据中的应用，表中所示的数据 2取自这些存储库；这些存储库通过开放许可证和公共 SPARQL端点分发数据。需要注意的是，这些端点可能并不总能反映图书馆的当前状况。[5] 数据抓取通过简单的脚本完成，[6] 例如附录 A。

资源的多样性 D 和丰富度 R 均已计算，同时还计算了多样性与丰富度之比，用以表明现有标签的使用效率。如表3所示，一些图书馆（如奥地利国家图书馆（AT）、芬兰国家图书馆（FI）和荷兰皇家图书馆（KB））采用的词汇表所包含的

[4]http://www.openarchives.org/pmh[5]例如，截至2022年3月，Europeana 的 SPARQL端点自2017年7月以来一直未更新。[6]部分存储库对下载设有超时限制。在这种情况下，需要使用分区查询来检索所有信息。

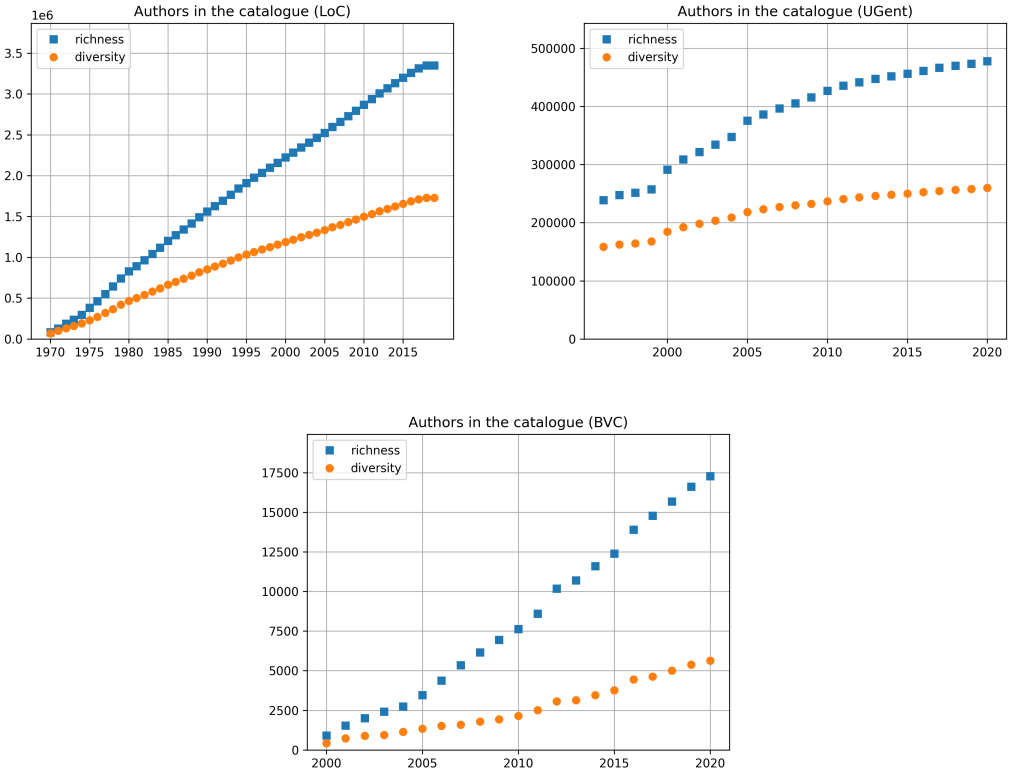

图5 目录中作者的累积数量及作者的香农多样性，按 MARC 记录进入目录的年份统计。

表2 图书馆发布的关联开放数据存储库。

机构	词汇表	URL
奥地利国家图书馆	EDM BIBFRAME RDA	labs.onb.ac.at/en/dataset/lod
西班牙国家图书馆	FRBR	datos.bne.es
塞万提斯虚拟图书馆	RDA	data.cervantesvirtual.com
法国国家图书馆	FRBR	data.bnf.fr
卢森堡国家图书馆	XML	data.bnl.lu
英国国家书目	BIBO	bnb.data.bl.uk
Europeana	EDM	pro.europeana.eu/page/sparql
德国国家图书馆	BIBFRAME	www.dnb.de/EN/lds
美国国会图书馆	BIBFRAME	id.loc.gov
芬兰国家图书馆	Schema.org BIBFRAME	data.nationallibrary.fi
荷兰皇家图书馆	Schema.org LRM	data.bibliotheken.nl

类和属性数量较少。相比之下，法国国家图书馆（BNF）和西班牙国家图书馆（BNE）使用更丰富的书目记录功能需求（FRBR）和资源描述与检索（RDA）词汇表来描述其资源。法国国家图书馆还采用了一套专有词表来描述创作者角色，其中包含 500 多个类别。由于这些类别的使用并不统一，这导致较低的 D/R 比率。英国国家书目（BNB）属于中间情况，因为它基本上采用了包含 33 个类和 88 个属性的 BIBO 词汇。

尽管类的多样性与



图 6 主题丰富度累积以及目录中主题的香农多样性指数。左：完整的主题标题。右：主题标题的细分项。注意丰富度所使用的特定尺度。

各个馆藏所采用的属性多样性之间存在中等程度的正相关（见图7），但有些图书馆的类粒度更细，而另一些则采用了种类更多的属性。

4 结论

多样性指数为数据馆藏中各组的多样性提供了互补的视角。在与丰富度相比，多样性对样本量的变化更为稳健，因为它对出现次数较少的类赋予的权重较低。

分析词汇内容时，词语的多样性会趋近于一个渐近值，该值取决于作品的作者和体裁。通过使用一个只涉及三个自由参数的简单模型对观测值进行外推，即可获得这一渐近值

资源类型 host	类 D	类 R	类 D/R	属性 D	属性 R	属性 D/R
AT	2.1	5	0.42	10.7	22	0.48
BNB	13.2	33	0.40	26.6	88	0.30
BNE	3.8	16	0.24	50.9	189	0.27
BNF	6.9	26	0.27	55.5	791	0.07
BVC	6.6	27	0.24	32.0	165	0.19
EU	5.1	11	0.46	37.1	115	0.32
FI	7.0	12	0.59	17.3	35	0.49
KB	3.9	12	0.32	14.6	23	0.64

表3 多样性 y D、丰富度 R 和关联开放数据中包含的资源的多样性-丰富度比率 D/R .

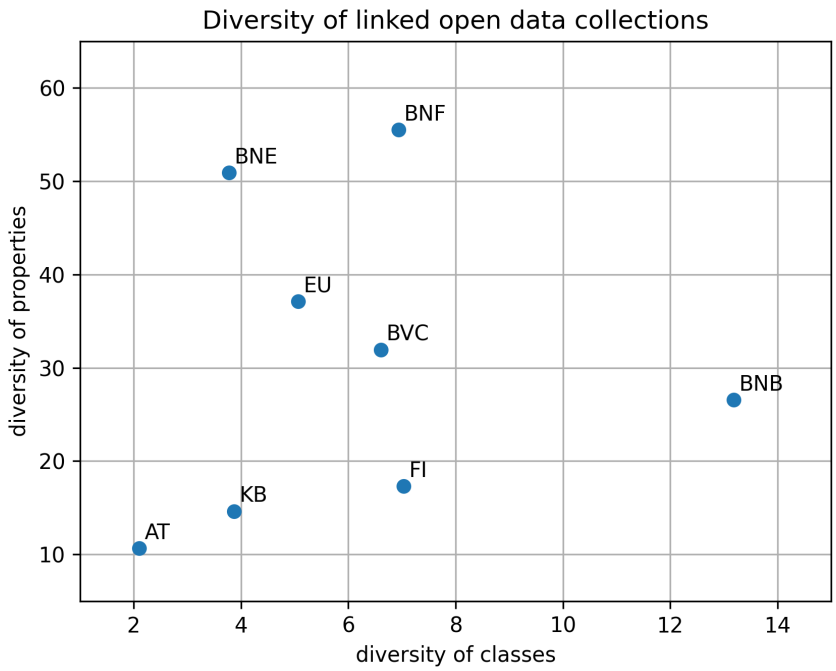

图7 图书馆在 *linked open data* 中发布的类和属性的香农多样性。

。这种外推对于样本大小具有稳定性。

就元数据而言，多样性指数可用于直观呈现趋势，例如创作者或主题覆盖范围的变化。多样性与丰富度之间的比率也有助于比较可用描述符（类和属性）在描述数字图书馆发布的语义数据（关联开放数据集）中的资源时的实际使用情况。

本文分析所用的 Python 脚本已作为开放获取软件发布于 [3]。

致谢。 感谢根特大学图书馆的 Frank Vande-pitte 和 Patrick Hochstenbach 热心协助我们了解该图书馆的编目记录。

附录A SPARQL查询

列表 1 用于检索 LOD存储库中所有类及每个类所含资源数量的查询。

```
选择 ? c l a s s (计数(? s ) 作为 ?计数 )其中
{? s a ? c l a s s}按组 ? c l a s s
```

列表 2 检索所链接的外部存储库的查询来自特定 LOD存储库的链接，以及指向每一个。

```
SELECT ?主机名 (COUNT(? s ) AS ?count )
WHERE{
    ? s owl : sameAs ?same .
     bind (
         s t r b e f o r e ( s t r a f t e r (
          s t r (? same ) , " // " ) , "/" )
          AS ? 主机名 )
}
按 ? 主机名 分组
```

参考文献

[1]Berners-Lee T（2006）关联数据。URL
h
t
t
ps://
www.w3.org/DesignIssues/LinkedData.html[2]
Berners-Lee T、Hendler J、Lassila O（2001）语义网。《科学美国人》284[3] Carrasco RC、Candela G、Such MM（2022）rccarrasco/dl diversity：首次发布。
https://doi.org/10.5281/zenodo.6389967,
URLhttps://doi.org/10.5281/zenodo.6389967[4]
Colwell RK、Coddington JA（1994）通过外推估算陆地生物多样性。《伦敦皇家学会哲学汇刊 B辑：生物科学》345(1311)：101–118。
https://doi.org/10.1098/rstb.1994.0091,
URLhttps://doi.org/10.1098/rstb.1994.0091[5]
Hill MO（1973）多样性与均匀度：一种统一的记法及其影响。《生态学》54(2)：427–432。
https://doi.org/10.2307/1934352[6]Jarvis S（2013）捕捉词汇多样性中的多样性。《语言学习》63(s1)：87–106。
htt
ps://doi.org/10.1111/j.1467-9922.2012.00739.x

[7] Kub´at M, Mili˘cka J (2013) 体裁中的词汇丰富度测量。《定量语言学杂志》20(4):339–349.https://doi.org/10.1080/09296174.2013.830552[8] Kyle K, Crossley SA, Jarvis S (2021) 使用直接判断评估词汇多样性指数的效度。《语言评估季刊》18(2):154–170.
https://doi.org/10.1080/15434303.2020.1844205[9] McCarthy PM, Jarvis S (2010) MTLD、vocd-d 和 HD-d：对词汇多样性评估高级方法的验证研究。《行为研究方法》42(2):381–392.
https://doi.org/10.3758/brm.42.2.381[10]McKee G, Malvern D, Richards B (2000) 使用专用软件测量词汇多样性。《文学与语言计算》15(3):323–338.https://doi.org/10.1093/llc/15.3.323[11] Piantadosi ST (2014) 自然语言中的齐夫词频定律：批判性综述与未来方向。《心理学公报与评论》21:1112–30.
https://doi.org/10.3758/s13423-014-0585-6[12] Roswell M, Dushoff J, Winfree R (2021) 测量物种多样性的概念指南。《Oikos》130(3):321–338.https://doi.org/10.1111/oik.07202,
U
R
L
h
t
t
p
s://
onlinelibrary.wiley.com/doi/abs/10.1111/oik.07202
[13] Shannon CE (1948) 通信的数学理论。《贝尔系统技术期刊》27(3):379–
4
2
3.
https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[14] Smith-Yoshimura K (2020) 迈向新一代元数据。
https://doi.org/https://doi.org/10.25333/rqgd-b343
[15]万维网联盟 (2013) RDF 的 SPARQL 查询语言。URL https://www.w3.org/TR/sparql11-overview/[16]万维网联盟 (2014) RDF 1.1 概念与抽象语法。URLhttps:

//www.w3.org/TR/rdf11-concepts/