DBpedia 和维基数据中的政治代表偏差：下游处理面临的挑战

Özgür Ka radeniz[1]*，Bettina Berendt[2,3,1]*，Sercan Kıyak[1], Stefan Mertens[1], Leen d'Haenens[1]

[1] 鲁汶大学，比利时；[2] 柏林工业大学，德国；[3] 魏岑鲍姆研究所，德国 * 通讯作者
{ozgur.karadeniz|bettina.berendt}@kuleuven.be

摘要：多样性搜索器最初是一种工具，旨在帮助分析新闻媒体文本中的多样性。它依托某种形式的自动化内容分析，因此建立在先前假设之上，并取决于与多样性和公平性相关的若干设计选择。其中一项设计选择是所使用的外部知识来源。本文探讨这些来源可能对内容分析结果产生的影响。我们比较多样性搜索器使用过的两个数据源——DBpedia 和维基数据——考察它们的本体覆盖范围和多样性，并说明这会对文本语料库的分析结果产生什么影响。我们介绍一项案例研究，考察 1990年至2020年间比利时政党在英语版 DBpedia、荷兰语版 DBpedia 和维基数据中的相对过度或不足表征，并重点指出在设计这项数据分析时需要作出的诸多决定、背后的假设，以及分析结果所带来的影响。尤其值得注意的是，我们发现英语版 DBpedia 对政治右翼存在惊人的过度表征。

1 引言

多样性搜索器最初是一种工具，旨在帮助分析新闻媒体文本中的多样性，如今已可供公众使用，也适用于公共图书馆。第 2 节介绍的多样性搜索器依托某种形式的自动化内容分析，因此建立在先前假设之上[1]，并取决于与多样性和公平性相关的若干设计选择。[2]接着在第 3 节中，我们研究底层数据源中的表征偏差，以及这些偏差可能对该工具所能识别的多样性程度产生的影响。

我们分析目前正在使用的两种替代方案——DBpedia 和维基数据——，考察它们的本体覆盖范围和多样性，并阐述这对文本语料库分析结果的影响。具体而言，我们介绍了一项案例研究，该研究考察了1990年至2020年间比利时政党的相对过度或不足表征，并指出开展这项数据分析所需作出的诸多研究设计决策、这些决策背后的假设，以及研究结果带来的影响。尤其值得注意的是，我们发现英语版 DBpedia 对政治右翼的表征严重过度。

这一结果为面向公共图书馆的数据和人工智能驱动的工具（如多样性搜索器）的设计提出了一个引人深思的挑战。该工具不仅能通过其基本功能发挥作用，或许更重要的是，也能通过公开呈现我们案例研究中的此类发现，帮助促进公平性和多样性，并让新闻用户意识到，多元声音新闻对于准确、公正地了解新闻报道所述现实至关重要。

2 多样性搜索器：文本分析工具

多样性搜索器是一款半自动化的文本分析和知识增强工具，旨在向用户展示新闻媒体文本的多样性程度。[3] 该工具是在跨学科项目 DIAMOND（多样性与信息媒体：多元公共辩论的新工具）[4] 中开发的，供媒体专业人士和媒体消费者使用，并与 iCandid[5] 集成。iCandid 是一个研究基础设施项目，为研究人员提供对多个（社交）媒体资源的集成访问。在项目开展期间，多样性搜索器得到进一步开发，也成为一款可供图书馆使用的工具。

该工具会分析上传的文本或文本集合。它会提取行动者（人物、机构和地缘政治实体），并提供有关这些实体的其他信息和交互

[1] 例如，参见 Kitchin, Rob：《数据革命：大数据、开放数据、数据基础设施及其影响》。伦敦：Sage，2014年。[2] Berendt, Bettina；Karadeniz, Özgür；Mertens, Stefan 等：《超越“平等”的公平性：多样性搜索器——检测并提升新闻媒体中社会政治行动者表征的工具》，载于《2021年万维网大会伴随论文集》，斯洛文尼亚卢布尔雅那，2021年，第202–212页。在线：<https://doi.org/10.1145/3442442.3452303>，访问日期：2022年8月25日。[3] 同上。[4] 关于DIAMOND，2018年1月7日，<https://soc.kuleuven.be/ims/diamond/diamond-about>，访问日期：2022年8月27日。[5] iCandid，《LIBIS研究基础设施概览》，2022年5月3日，<https://bib.kuleuven.be/english/libis/projects#icandid>，访问日期：2022年8月25日。

机会，涵盖个体行动者及其在文本中的出现情况。这些内容包括基础统计数据（例如不同行动者的频率），以及对“多样性”这一复杂概念的数值评估。

2.1 文本中（行动者）多样性的定量测量

我们的核心假设之一是，媒体多样性是民主健全运作和公众舆论充分知情的重要前提。多样性被视为信息质量的重要组成部分。为此，我们借鉴了 Stirling（2007）[6] 颇具影响力的提议：提出一种定量测量方法，该方法建立在对多个学科多样性文献的元分析之上。Stirling 指出了三个能够提升多样性的组成部分：

> - **多样性类别：种类丰富度**：存在许多不同实体，例如来自不同背景的多个人；- **差异**：实体彼此不同，涵盖范围广泛；- **均衡度**：不同实体的分布均匀。这一组成部分是公平性文献中“……平等”概念的一种简单形式。

基于这三个组成部分，Stirling 将多样性定义为Δ

$$\Delta = \sum_{i,j(i \neq j)} (d_{ij})^{\alpha} \cdot (p_i \cdot p_j)^{\beta}$$

其中，多样性类别：种类丰富度由领域（例如文本、人口等）中所有实体 i、j ∈E 集合的基数表示；均衡度由频率 p_i 表示（分布越均匀，这一乘数越大）；差异则由距离或不相似度的测量值 d(.,.) 表示。此外，参数 α 和β 可用于对均衡度或差异的重要性进行加权。

因此，新闻媒体中的多样性意味着有大量新闻报道可供选择（多样性类别：种类丰富度），且这些报道表达不同观点（差异），同时所有观点都得到同等表征（均衡度）。正如 Ranaivoson 指出的，使用这一公式时，差异常常成为障碍，因为其衡量依据是一些强假设。[7] 因此，该工具着眼于媒体内容多样性的一个特定方面：行动者多样性。

2.2 多样性搜索器处理流程概述

在本小节中，我们简要概述多样性搜索器的处理过程。关于详情、用户界面和完整示例，请参见（Berendt 等人，2021年）。[8]

---

[6] Stirling, Andy：科学、技术与社会多样性分析的一般框架，载于：皇家学会界面期刊 4 (15)，2007年8月22日，第 707–719 页。在线：<https://doi.org/10.1098/rsif.2007.0213>. [7] Ranaivoson, Heritiana：使用斯特林模型衡量文化多样性，载于：统计学新技术与新方法，2013年，第 10 页。在线：<https://ec.europa.eu/eurostat/cros/content/measuring-cultural-diversity-stirling-model-heritiana-_ranaivoson_en>。[8] Berendt, Bettina；Karadeniz, Özgür；Mertens, Stefan 等：《超越“平等”的公平性：多样性搜索器作为检测和提升新闻中社会政治行动者表征的工具

我们研究社会政治行动者的多样性，并广义地将其理解为具有社会政治能动性的三类行动者：人物、组织和地缘政治实体。我们的设计选择受到一种社会学理解的启发[9]，这种理解关注两个社会行动者在多大程度上以及以何种方式存在差异。

随后，我们构建了一个知识检索和文本处理管道，该管道利用底层知识来源本体中的数据。最初，我们使用的是英语版 DBpedia。多样性搜索器分两个步骤进行分析。第一步是结合命名实体识别（NER）、命名实体链接（NEL）以及区分大小写和不区分大小写的文本与词元匹配，从文本中检测行动者。[10] 对于 NEL，我们使用 DBpedia Spotlight，这是一种支持多种语言的 NEL 模型，可将实体链接到 DBpedia。[11] NER 和模式匹配任务使用第三方工具 spaCy，这是一款免费提供的自然语言处理库及语言模型。[12] 当媒体研究领域的同事发现 Spotlight 对佛兰德政治人物的识别效果不佳时，我们便在管道中加入了模式匹配功能。这一功能的加入也为命名实体以及“难民”“罪犯”等未命名实体的用户定义识别规则铺平了道路。

该应用程序包含语料库上传功能。用户可以通过此功能上传从 Belga Press 导出的 XML 文件[13] 并对其进行批量处理，还可以将结果导出为 Excel 文件，以便在 Excel 或 SPSS 中进行进一步分析。

虽然识别行动者足以处理均衡度和多样性类别：种类丰富度，但差异还需要关于已识别实体的更多信息，因此需要知识检索与丰富阶段。由于 DBpedia 中既包含与我们的任务相关的信息，也包含无关信息，因此分析还需要筛选我们从 DBpedia 获取的谓词-对象对。为此，我们开发了自己的本体，其中包含具有社会政治相关性的行动者及其特征。于是，DBpedia 中的图数据会根据我们的本体转换为一组“feature_name”：“feature_value”键值对，并插入该资源的本地记录中，以供后续计算使用。

知识检索与本体映射会处理所有链接到根资源的资源，前提是它们也属于三种行动者类型。此决定基于这样一个事实：与根实体相链接的某些类型实体的属性，也包含有关根实体的重要信息。因此，多样性搜索器不会在检索后就停止，例如

媒体，载于：《2021年万维网大会伴随论文集》，斯洛文尼亚卢布尔雅那，2021年，第 202–212 页。在线：<https://doi.org/10.1145/3442442.3452303>，检索日期：2022年8月25日。 9 皮埃尔·布迪厄：《社会空间与群体的生成》。Theory and Society 14(6)，1985年，第 723–744 页。10 NER：自然语言处理（NLP）的一项子任务，即将文本中的命名实体归入预定义类别；NEL：自然语言处理的一项子任务，即将文本中的命名实体链接到远程知识库中的资源。11 Mendes, Pablo N.；Jakob, Max；Andrés García-Silva 等：《DBpedia Spotlight：揭示文档网络》，载于：第七届国际语义系统会议论文集——I-Semantics ’11，奥地利格拉茨，2011年，第 1–8 页。在线：<https://doi.org/10.1145/2063518.2063519>，检索日期：2022年8月25日。12 Matthew Honnibal；Ines Montani：《spaCy 2：利用 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解》，2017年。<https://spacy.io/>，检索日期：2022年8月29日。13 Belga Press（belgapress.be）是一个在线新闻数据库，收录了大量国际及比利时新闻来源，记者和媒体研究人员可在其中搜索并导出语料库。具体来源及其 XML 格式与我们的目标用户相关，但该架构很容易改为可配置。

检索政治人物所属的政党：它还会尝试检索该政党的属性，例如意识形态，因为不同政党（多样性类别：种类丰富度）的意识形态可能存在交集，从而影响差异。同样，找到行动者所属的国家后，还会进一步查询该国家的属性，例如欧盟成员资格或政府类型。

尽管通过进一步补充行动者的背景信息，这些关联实体的信息有助于丰富内容，但整合这些信息也有一个缺点：这些资源中的错误和偏差也会一并引入。例如，本地知识库中关于“希拉里·克林顿”“乔·拜登”和“唐纳德·特朗普”的数据，也会受到“美国”条目数据质量的影响。同样，与特定意识形态政党的表征相关的任何系统性政治偏见，都会影响工具对根政治人物的表征，即使这些人物被正确识别也是如此。

3 底层本体在偏见与多样性方面的作用：比较 DBpedia 与维基数据

多样性搜索器这类任务在很大程度上依赖外部知识源。选择使用哪些知识源时，需要比较不同选项。问题在于，如何以正式、自动化的方式进行比较，同时支持应用的目标。我们选择了一种方法，该方法基于：(a) 对可用外部知识源的一系列观察和假设；(b) 将这些假设落实为计算程序，从资源描述框架结构中提取关系表示；以及 (c) 针对所获数据提出的特定应用问题。一方面，评估标准基于 Zaveri等人提出的用于确定关联开放数据（LOD）数据质量的综合框架；另一方面，也参考了关于衡量语言（模型）偏见的文献，并结合我们关于行动者多样性的问题对这些标准进行了重新诠释。[14] [15]

例如，即使一篇新闻媒体文章具有很高的多样性（比如，其中所呈现行动者的政治派别多样），如果实体识别过程依赖的知识来源只包含某一政治派别的行动者，那么在多样性计算中只会处理这些行动者，最终得到的多样性值也会很低。因此，理想的知识来源应具有较高的覆盖度、正确性和时效性，并且多样性较高。

因此，我们开始质疑最初选择 DBpedia 是否有助于开展相关讨论：与所研究国家或地区的语言（如适用）相比，选择英语版 DBpedia 会带来什么后果？（一般而言，地方版 DBpedia 对当地行动者的覆盖度更高。）我们也开始怀疑，采用更严格质量控制的维基数据[16] 是否能为多样性搜索器提供更好的基础。

---

[14] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer：《关联数据质量评估：一项综述》。语义网 7(1)：63-93（2016年）[15] Delobelle, Pieter；Tokpo, Ewoenam；Calders, Toon 等：《使用偏置标尺测量公平性：预训练语言模型偏见度量指标的比较研究》，载于：《2022年计算语言学协会北美分会年会论文集：人类语言技术》，美国西雅图，2022年，第1693–1706页。在线：<https://doi.org/10.18653/v1/2022.naacl-main.122>，检索日期：2022年8月25日。[16] Piscopo, Alessandro；Simperl, Elena：《谈论维基数据质量时，我们在谈论什么：一项文献综述》，载于：《第十五届开放国际研讨会论文集》，瑞典舍夫德，2019年，第1–11页。在线：<https://doi.org/10.1145/3306446.3340822>，检索日期：2022年8月29日。

这些问题给工具设计带来了两类挑战。第一类与实际操作和实现有关，第二类则是概念层面的。在关联开放数据时代，知识来源的选择看起来可能是一项模块化、因而易于配置的设计特性。事实上，在多样性搜索器这类工具中，可以实现多种选择，包括在这些知识来源之间进行灵活选择和／或由用户决定的选择[17]。然而，这需要先对这些数据集开展研究，并调整提取算法，因为谓词名称和编码约定各不相同（例如 DBpedia 与维基数据之间），而且即使在 DBpedia 内部，谓词可能具有普遍适用性这一特点也并非总能得到利用，不同语言版本还会倾向于使用不同的谓词，其语法和语义相关但并不完全相同。例如，在英文 DBpedia 中，政治人物任职期限的时间维度编码方式被广泛采用，但其他知识来源并非如此；在这些来源中，通常会标明担任职位。此类差异可能是因为某一语言版 DBpedia 的作者社群成员，沿用了自己社群中其他人呈现信息的风格。此外，在每种情况下，工具开发者都有理由选择保持统一（提取相同的谓词），或依据其他标准选择谓词。我们采用了后一种做法，从语义高度相关的谓词中，选取覆盖度最高的谓词。

3.1 基本统计

我们首先选取英文 DBpedia 和维基数据作为两个外部知识源；此方法也很容易推广到其他来源。作为关联开放数据中最大的两个来源，DBpedia 和维基数据都能够为多样性搜索器识别的实体类型提供相关信息。[18] [19] [20] 虽然两者都包含基于维基百科、可供机器读取且免费开放的数据，但它们的本体和构建方式有所不同。

DBpedia 是两者中较早的一个，其定义是“一个由众包推动的社区项目，旨在从各种维基媒体项目创建的信息中提取结构化内容。”[21] DBpedia 从维基百科页面（主要是信息框）中提取信息并将其表示为 RDF，同时以半自动的方式根据页面为其资源创建唯一的（依语言而异的）URI

[17] 也可以组合多个知识来源，但这需要做出更多选择，尤其是在两个来源的信息相互冲突时。因此，这个问题将留待未来研究。[18] Auer, Sören；Bizer, Christian；Kobilarov, Georgi 等：《DBpedia：开放数据网的核心》，载于：Aberer, Karl；Choi, Key-Sun；Noy, Natasha 等（编）：《语义网》，第 4825 卷，柏林、海德堡，2007 年（《计算机科学讲义》），第 722–735 页。在线：< 10 1007 978 3 540 76298 0_52> https://doi.org/。/ -- - - Lehmann, Jens；Isele, [19] Robert；Jakob, Max 等：《DBpedia——从维基百科抽取的大规模多语言知识库》，载于《语义网期刊》6（2），2015 年，第 167–195 页。[20] Vrandečić, Denny；Krötzsch, Markus：《维基数据：一个免费的协作式知识库》。载于：《ACM通讯》57（10），2014 年，第 78–85 页。在线：< http://dx.doi.org/10.1145/2629489>。[21] 关于 DBpedia，DBpedia，（无日期）。

的方式。[22] 通过这种方式，它构建了一个开放的语言图谱，与语义网良好集成，并受到关联开放数据社区的欢迎。[23]

维基数据与 DBpedia 在许多方面有所不同。首先，两者与维基百科的关系不同。维基数据旨在为其他所有维基媒体项目提供结构化数据，并包含维基百科未涵盖的实体。然而，DBpedia 严格受限于维基百科上可获取的信息。其次，维基数据由一个规模庞大且开放的贡献者社区编辑和维护，他们通过各种维基专题核查事实并丰富数据。[24]DBpedia 基于一个封闭的社区运作，并且更依赖维基百科上信息的正确性。第三，维基数据集不依赖语言，由一个大型图谱组成，可使用任何可用的语言标签或描述来呈现。与依赖语言的 DBpedia 数据集不同，维基数据中的属性和实体分别以 P 编号和 Q 编号作为 URI。第四，得益于其资源描述框架结构，维基数据允许为陈述添加限定符和参考文献，而这些额外信息有助于各种自然语言处理任务。最后，尽管两者都提供 SPARQL API，维基数据的限制更多，因此不太适合提取大量数据。[25]

DBpedia 和维基数据提供的信息量也有所不同。DBpedia 英语版本的最新快照包含超过 8.5 亿个三元组，描述了约 540 万个实体。[26] 另一方面，维基数据提供约 13.6 亿个三元组，描述了约 9700 万个实体。[27] [28] 因此，DBpedia 中每个实体的平均陈述数约为 157 条，而维基数据中约为 14 条。由此可见，正如先前研究所指出的，维基数据描述的实体更多，而 DBpedia 对每个实体包含的陈述更多。[29]DBpedia 以多种语义语言提供有关其实体的各类陈述，而维基数据提供的资源更多，其中包含更结构化、更精简的信息，能够

---

[22] Hofer, Marvin；Hellmann, Sebastian；Dojchinovski, Milan 等：《DBpedia全新发布周期：提升知识抽取工作流的敏捷性与效率》，载于：Blomqvist, Eva；Groth, Paul；Boer, Victor de 等（编）：《语义系统：知识图谱时代》，第12378卷，Cham，2020年（《计算机科学讲义》），第1–18页。在线：<https://doi.org/10.1007/978-3-030-59833-4_1>，检索日期：2022年2月21日。 [23] Abián, D.；Guerra, F.；Martínez-Romanos, J. 等：《Wikidata 与 DBpedia：一项比较研究》，载于：Szymański, Julian；Velegrakis, Yannis（编）：《结构化数据源上的语义关键词搜索》，第10546卷，Cham，2018年（《计算机科学讲义》），第142–154页。在线：<https://doi.org/10.1007/978-3-319-74497-1_14>，检索日期：2022年8月25日。 [24] 维基数据：统计，<https://www.wikidata.org/wiki/Wikidata:Statistics>，维基数据，（无日期），检索日期：2022年3月1日。[25] API 的当前限制可通过以下链接查看：<https://www.dbpedia.org/resources/sparql/>；<https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual Holze, Julia：《DBpedia 快照 2021 12 09 02 2022 - 发布公告》，<DBpedia博客，. .，https://www.dbpedia.org/blog/snapshot-2021-12-release/，检索日期：. . . [26] [27]维基数据仪表板 / 维基数据数据模型陈述，Grafana，（无日期），<https://grafana.wikimedia.org/d/000000175/wikidata-datamodel-statements?orgId=1&refresh=30m>，检索日期：2022年3月1日。[28]维基数据：统计，（无日期）[29] Abián 等：《Wikidata 与 DBpedia：一项比较研究》，2018年 28 08

在某些特定用例中发挥更高的效用。对文献研究结果的这一概述也与我们自己的结果相符，接下来将介绍这些结果。[30]

3.2 对政治人物领域覆盖度的简要探究

在首次数据探索中，鉴于英文 DBpedia 和维基数据与多样性搜索器的相关性，我们重点考察了其中政治人物覆盖度的示例。我们提出了以下 SPARQL 查询：

查询	英语版 DBpedia	维基数据
比利时众议院议员 比利时众议院 [31]	143	2996
佛兰德议会议员 议会 [32]	21	464
曾在.....任职的政治人物 美国众议院的 众议员 [33]	14886	11,160

表1. 用于比较英语版 DBpedia 和维基数据中政治人物覆盖度的三个查询。

查看第二个查询的结果后，我们注意到，所提供的佛兰德政治人物中有 15 位（约占四分之三）属于同一个政党（N-VA，即 Nieuw-Vlaamse Alliantie、新佛兰德联盟）。尽管 N-VA 目前是佛兰德议会中最大的政党，拥有 35 个席位

30 Färber, Michael、Frederic Bartscherer、Carsten Menne 和 Achim Rettinger。《DBpedia、Freebase、OpenCyc、维基数据和 YAGO 的关联数据质量》。Amrapali Zaveri、Dimitris Kontokostas、Sebastian Hellmann、Jürgen Umbrich、Amrapali Zaveri、Dimitris Kontokostas、Sebastian Hellmann 和 Jürgen Umbrich 编。语义网 9，第1期（2017年11月30日）：77–129。https://doi.org/10 8 31 3233 170275 46 . /SW- . pp.- DBpedia 页面链接：https://dbpedia.org/page/Category:Members_of_the_Chamber_of_Representatives_(Belgium)，维基数据查询链接：https://w.wiki/4DmS。不过，我们还应指出，荷兰语-佛兰德语 DBpedia 中比利时众议院成员的同一名单包含 701 个资源：http://nl.dbpedia.org/page/Kamer_van_volksvertegenwoordigers。访问日期：01/10/2021。接下来的章节将进一步探讨这三个数据库各自的表征问题。32 DBpedia 查询链接：https://tinyurl.com/4uwbycwu，维基数据查询链接：https://w.wiki/3YG4。访问日期：01/10/2021。33 DBpedia 查询链接：https://tinyurl.com/e3y2b8kh，访问日期和维基数据查询链接：https://w.wiki/4DnP。访问日期：01/10/2021。

在总计 124 个席位中，与其他弗拉芒政党相比，该党的成员在英文 DBpedia 中所占比例过高。数据还显示，DBpedia 中存在一些分类错误。[34] [35]

这些结果表明，英文 DBpedia 提供的实体可能较少（尤其是在非英语语境中），而且在偏见和质量方面似乎存在问题。我们的观察与关于这两个信息来源的现有比较研究一致。[36]

3.3 深入考察政党的表征

3.3.1 研究动机：右翼人士是否已经接管维基百科/DBpedia？以及这为何表明我们需要审视多样性和偏见

为了更好地理解 DBpedia 中可能存在的表征偏差，我们进一步仔细考察了这类偏差，以及它们对多样性搜索器所理解的多样性可能意味着什么。具体而言，我们探讨了如何捕捉由“过度表征”造成的偏差，以及在多样性搜索的语境中，首先应如何定义“源本体中的过度呈现”。鉴于政治分析总是需要结合其历史背景来理解，我们还探讨了：对于一个当前状态不断演变的知识来源而言，对历史事实（也）过度呈现可能意味着什么。

这些问题也促使我们探究 DBpedia、维基数据等作为文化记忆基础设施的作用。就本文考察的案例——按政党对政治人物进行表征——而言，可以认为，议会记录、政党记录等可能是规范、准确且全面的来源，因此使用 DBpedia/维基数据主要是出于务实和节省成本的考虑；但多样性搜索器所需的表征及输入信息当然远不止政治人物及其政党隶属关系，因此，尽管众包来源存在种种缺陷，它们往往仍是文化记忆的唯一形式。由此，这项研究既关注选择来源时务实考量所带来的后果，也研究来源本身。

3.3.2 方法

接下来的分析分为六个阶段。

1. 我们重点考察了“政党隶属关系”和“政党的政治倾向”这两项**属性**，并以此作为评估偏见和多样性的依据。我们将分析范围限定为“来自比利时的政治人物”，因为我们很早就意识到，即使要理解纯粹的数值现象，也需要扎实的领域知识，而我们的团队成员正是这一领域的专家。

---

[34] 政党成员信息根据前文提到的 DBpedia 查询（https://tinyurl.com/4uwbycwu）手动收集。访问日期：01/10/2021。 [35] 例如，DBpedia 将妇女平等党（纽约）视为一种“人物”，并将其列为政治人物：https://dbpedia.org/page/Women%27s_Equality_Party_(New_York) [36] Abián 等：《Wikidata 与 DBpedia：一项比较研究》，2018

2. 根据**零假设**，如果某个政党在特定现实情境中的相对人数，与知识来源中被表示为实体的该党政治人物的相对人数相等，那么该知识来源对政治格局的呈现就是无偏的[37] 。我们选择使用相对人数，是因为从一开始就很清楚，我们的知识来源并非规范性且全面的公共记录；此外，我们认为“拥有维基百科页面”（等等）是提升个人及其所属政党公众可见度的重要因素。

这个一般性问题只有在特定时间点上才有意义，因为基准会不断变化，至少会随着改变政治代表性的选举而变化。

3. **基准**随后相对容易定义：政党 Y 在给定政治机构中各个时间点 T 的得票份额或席位数量。我们首先将重点放在国家议会，即众议院（*Kamer vanvolksvertegenwoordigers*，下文简称 *KVV*），并采用每个立法会期开始时的席位数量。我们还考察了地区（佛兰芒）议会（*Vlaams parlement, VP*），其选举时间点与国家层级的选举时间点并不一致。

T这些数据取自*Wikipedia, de vrije encyclopedie*（2022a、2022b）。[38] [39]

4. **本体中的表征**较难定义。人们可能希望将 T 时刻政党的得票份额（第3步）与 T 时刻本体中所表征的内容进行比较。我们选择将其与本体“当前所表征、但作为 T 时刻的相关内容”进行比较。原因有二。首先，我们希望研究自1990年以来的历史演变，而维基百科直到2001年初才开始运营，此后还经历了一段发展时期，尤其是其非英语版本（例如，荷兰语版本创建于2001年）。维基数据于2012年10月上线。其次，我们旨在将这些本体也作为文化记忆的基础设施来研究，这使问题略有不同：当前用户是如何透过本体的视角来看待政治格局（如1990年代等）的?

使用当时撰写本文时的本体，通过 SPARQL 查询创建了“当前具有代表性”的行动者和政党的 CSV 文件。[40] 在接下来的第 5 步中进行 “ 在 T 时刻相关 ” 的筛选。

---

37 对于源自维基百科的 DBpedia，这对应于“拥有维基百科页面的人”。38 Wikipedia-bijdragers: Kamer van volksvertegenwoordigers — Wikipedia, de vrije encyclopedie，2022年。在线：<https://nl.wikipedia.org/w/index.php?title=Kamer_van_volksvertegenwoordigers&oldid=61849646>，检索日期：2022年8月25日。39 Wikipedia-bijdragers: Vlaams Parlement — Wikipedia, de vrije encyclopedie，2022年。在线：<https://nl.wikipedia.org/w/index.php?title=Vlaams_Parlement&oldid=61662474>，检索日期：2022年8月25日。40 按照多样性搜索器针对在线本体的查询方式，我们最初设计查询，旨在获取比利时政治人物的信息，以及有关其政党和国家的背景信息。关于所查询属性的详细信息及其设计依据见（Berendt 等人，2021年），也可通过超链接查询查看。由于与研究问题无关（例如政治人物的性别或国家信息），案例研究分析未使用其中部分信息。为适应差异，有必要对查询作出一些调整并采用变通方法

5. **计算政党代表份额**。那么，我们如何从“本体中存在政治人物 X”得出“政党 Y 覆盖了 T 时刻表征的 Z%”？

5a. 我们首先将分析范围限定为在 T 时仍活跃的政治人物。我们通过将本体中表示的所有活动时期合并，得出每位行动者最长的活动时期；若未明确注明结束日期，则以“今天”为截止日期，或在适用时以行动者的死亡日期为截止日期；少数情况下，则采用广为人知的[41] 日期作为其政治生涯结束日期。我们选取了在 T = 1 1990年、1996年、2000年、2005年、2011年、2015年和2020年1月仍活跃的政治人物。（之所以未遵循每隔 5 年取样，是为了捕捉 1995 年和 2010 年举行的大选所产生的影响。）

5b. 随后，我们对每位行动者的政党隶属关系列表进行了预处理（该列表通常既冗长又多样）。首先，将曾在一个或多个时间点更改名称的政党统一归为一个政党，并以其当前缩写作为标识。其次，由于占比很低，我们将小型政党归为“无关政党”“外国政党”等类别，不再进一步分析。这些映射由我们团队中的领域专家创建。

5c. 第三，这仍然遗漏了许多在职业生涯不同阶段属于不同政党的政治人物。乍看之下，这似乎是时态数据库中的“有效时间”问题，而所有最先进的数据库都能处理这一问题。有效时间是指数据库事实在所建模现实中有效的时间段。原则上，识别出媒体文本中某个行动者的工具[42] 可以查询行动者在相关时间 T 属于哪个政党 P，并在该语境下将其视为 P 的代表。然而，在评估文本时，这个时间段究竟指什么？首先，哪些时间段应该算数：文章撰写的时间、文章描述的时间，还是文章被阅读的时间？其次，如何识别并界定文本所谈论的时间段？第三，某个行动者的（事实或感知上的）“身份”难道不是由其完整经历塑造的吗？因此，多样性搜索器并未尝试从数据库事实中识别并处理有效时间，而是默认采取“立足当下”的视角，从而也模仿了人类用户。

但是，那个被模仿的人类用户会“看到”什么呢？从当下来看，这位政治人物可能被认为只让其中一个政党获得可见度，也可能让所有政党都获得可见度。如果他们让所有政党都获得可见度，可能是平等对待，也可能是其中一个党派关系“遮蔽”了其他关系。甚至可能是，这些历史变迁让人们只关注

不同本体之间的差异、本体的数据质量，以及结果最大数量方面的技术限制。第 6 节讨论了这些务实选择的局限性。查询可在以下地址找到：https://tinyurl.com/ys2bnyy7,https://tinyurl.com/ys2bnyy7，https://tinyurl.com/4ejp5jne，https://tinyurl.com/33fej4aw，https://tinyurl.com/2a3ph5bt（英语DBpedia）；https://w.wiki/56Qc，https://w.wiki/53Lb，https://w.wiki/53ap，https://w.wiki/53gw （维基数据）；https://tinyurl.com/2bjxp2f7，https://tinyurl.com/y5fzpu6p，https://tinyurl.com/56xkpy6m，https://tinyurl.com/38krrb57https://tinyurl.com/38krrb57（荷兰语 DBpedia）。访问日期（依次为）：27 2022年，12 052022，27 05 2022，12 04 2022，27 05 2022，12 04 2022，13 04 04 2022，13 04 2022，26 04 2022，13 04 2022，12 04 2022，12 04 2022。用于绕过查询限制的 Python 脚本见补充文档。[41]我们使用维基百科作为知识来源来确定这些案例。[42]关于多样性搜索器如何识别和处理行动者的简要说明，请参见第 2 节。

此人，而忽略了他们也属于某个政党。对于某位政治人物在某个特定时间点而言，哪种情况适用，很可能在很大程度上取决于这位政治人物、该时间点、政党及其周围所形成的政治行动和叙事[43]，用户，也可能取决于其他因素。由于缺乏确定这些情况分别适用于何种情形所需的心理学实证研究，我们选择通过表征的上下限来界定各种可能性：

在时间 T，政党至少通过所有在整个职业生涯中仅属于该政党的政治人物[44]而变得可见。这个人数定义了其在 T 时可见度的下界，因为这假设曾属于多个政党的政治人物不会为其中任何一个政党带来可见度。政党至多通过所有职业生涯中曾在任何时候属于该政党的政治人物而变得可见。这个人数定义了该政党在 T 时可见度的上界，因为这假设曾属于多个政党的政治人物会为所有这些政党带来全部可见度。不同的多党籍政治人物为其各个政党带来更多或更少可见度的所有情形，都介于这两个界限之间。

然后将这些界限与基准份额进行比较。如果下界高于基准，该政党肯定是过度代表。（例如：某政党获得了20%的选票，但本体中的所有政治人物里有30%是该党的纯粹成员，而且可能还有其他人曾在某个时候属于该党。）如果上界低于基准，则该政党肯定代表不足。（例如：某政党获得了20%的选票，但本体中的所有政治人物里只有10%曾属于该党……也可能还属于其他政党。）如果基准介于下界和上界之间，就无法明确判断该政党是过度代表还是代表不足。

步骤 5 是在 MySQL 关系数据库中进行后处理和查询，并将步骤 4 生成的 CSV 文件导入其中。

6. **S研究可能存在的过度／不足表征及其随时间变化的趋势**。该分析采用描述性方法，并借助下一节展示的可视化结果。原则上，我们应从统计角度明确“较高”和“较低”（既要考虑统计显著性，也要考虑效应量）。然而，鉴于行动者数量较少，且目前尚不了解其分布情况，我们仅进行描述性统计。

3.3.3 结果与阐释

结果汇总于图1中。该图突显了数据中的趋势以及知识来源之间的对比：政党按从左派到右派的顺序排列，曲线则随时间始终位于基线之上或之下。更高分辨率或可交互的在线版本中可查看详情。[45] 粗线位于细线之上表示该政党被过度表征，位于细线之下则表示该政党表征不足；从这个意义上说，其所属的政治倾向也同样如此。英语版 DBpedia 在2005年之前收录的行动者不足20名，因此难以解读过度表征和表征不足的情况。荷兰语版 DBpedia 中表征的政治人物数量在最后一轮地区选举和全国选举（2019年）之后急剧下降，这使得这张图也比其他图更难解读。

[43]例如右派政党与极右翼政党之间的 rapprochement 可能导致二者的界限变得模糊[44]我们不考虑成员身份属于选票数微不足道的“无关”小党的情况。[45] http://www.berendt.de/DIAMOND/，检索日期：2022年12月27日。

版 DBpedia 中表征的政治人物数量在最后一轮地区选举和全国选举（2019年）之后急剧下降，因此与其他图相比，这张图也更难解读。

这些结果不仅证实了我们最初的非正式观察：英语版 DBpedia 中右派政党（尤其是 N-VA）的表征过度，而且这种趋势还在随时间加剧。（这些年间，N-VA 的普选得票份额有所上升，但 DBpedia 的增长显然超过了基准增长。）荷兰语版 DBpedia 似乎存在不同的偏差：尽管总体上与基准相当接近，但这一本体似乎对主要中间派政党（基督教民主与弗拉芒党）的表征过度。相比之下，维基数据相当准确地呈现了各政党在国家议会中的席位占比。法语瓦隆政党在荷兰语版 DBpedia 中代表不足（鉴于其语言侧重，这可以理解）。此前的国际研究已指出媒体报道对右派和中间派政党的过度表征，例如媒体报道对2017[46] 英国选举中间派政党的偏好，以及尽管美国有人高呼遭到审查，社交媒体上仍存在对右派的过度表征。[47]

此处插入图1

荷兰语版 DBpedia 中政治中间派似乎代表过度，这可能是语言造成的假象：荷兰语 DBpedia 更关注弗拉芒政治人物，而非讲法语的瓦隆（或布鲁塞尔、德语社群）政治人物；对该本体中政治人物页面的非正式检查表明，其中收录了许多地区和地方政治人物。因此，图2以佛兰德议会的席位占比（该议会于1995年首次举行直接选举）为基准，其他方面则与图1类似。政党集合是图1所示集合的子集，因为只有弗拉芒政党能够进入佛兰德议会（而国家议会也有来自瓦隆、布鲁塞尔和德语社群政党的代表）。图2显示，在荷兰语版 DBpedia 中，从极左到中间派的政党表征实际上与这些政党在地区议会中的席位占比相当接近；而右派和极右往往代表不足，尤其是在后期——当时它们在选举中，特别是在弗拉芒选举中取得了巨大成功。

图2插入于此处

46 David Deacon、John Downey、David Smith、James Stanyer 和 Dominic Wring。《2017年大选的全国新闻媒体报道》。传播与文化研究中心，拉夫堡大学报告4：2017年5月5日—6月7日 传播与文化研究中心，拉夫堡大学 在线版：<https://blog.lboro.ac.uk/crcc/wp-content/uploads/sites/23/2017/06/media-coverage-of-the-2017-general-election-campaign-report-4.pdf>[47]Mark Scott：尽管有人高呼审查，保守派仍主导社交媒体，POLITICO，2020年10月26日， <https://www.politico.com/news/2020/10/26/censorship-conservatives-social-media- 432643>，检索日期：2022年8月28日。

鉴于小党代表在基线和本体中的占比都很低，且我们的研究重点是弗拉芒政党，因此不再进一步考察小党或瓦隆政党的表征。

3.3.4 对多样性搜索器的启示与未来工作

政治学文献此前曾指出，政治右翼善于运用社交媒体，这似乎是一种国际现象。至于“中间派表征过度”，这可能是荷兰语资源侧重荷兰语（佛兰德）政治人物以及区域层级多数派所造成的假象。将结果与佛兰德议会的基准进行比较后，过度表征现象便消失了，这也支持了这一解释。

在多样性搜索器的语境下，这类偏差为何重要？该工具只能计算其能够识别的实体之间的多样性。设想（a）一段均衡度较高的文本，其中有若干左翼行动者和若干来自某个右翼政党的行动者；以及（b）一段只包含右翼政党和极右翼政党行动者的文本。假设这两段文本在其他方面都相同，那么“真实情况”应是（a）比（b）更多样。然而，如果数据来自一个偏向政治右翼的本体，该工具在（a）中只会（或主要会）识别出那个右翼政党的行动者，而在（b）中会识别出来自这两个政党的大多数行动者，从而使文本（b）的计算所得多样性得分更高，而文本（a）的得分更低——甚至可能为零。这一结果与本体知识的一种假象有关（人类身上也能观察到）：人们会在自己熟悉的领域感知到更多差异和多样性。

案例研究还揭示了多样性、公平性与偏差之间复杂且多重的依存关系。“政党”是用于计算已识别实体之间 Stirling差异的特征之一。因此，文本计算得出的多样性会受到这一特征的影响。一般而言，这一特征不同取值（即不同政党）的均衡度越高，多样性就越大。然而，如果这些取值的占比偏离基准过多，那么该知识来源就存在偏见——这反过来可能影响文本多样性的计算，从而引入不公平性。

相比之下，维基数据对比利时议会中各政党的表征相当准确。在未来的研究中，还应调查其他相关领域的数据质量（例如其他国家的政党体系或其他社会政治相关主题），以便提供基于证据的理由，决定在多样性搜索器这类工具中选用哪种本体。

4 局限性、结论与展望

正如 Leorke 等人所指出的，向数字经济和智慧城市的转型改变了公共图书馆在社会中的角色。[48] 如今，公共图书馆越来越被视为并被期待发挥“枢纽”或“平台”的作用，“将人们与信息、服务

[48] Leorke, Dale；Wyatt, Danielle；McQuire, Scott：“不仅仅是一座图书馆”：智慧城市中的公共图书馆，载于：《城市、文化与社会》15，2018年12月，S. 37–44。在线版：<https://doi.org/10.1016/j.ccs.2018.05.002>。

以及彼此联系起来”。[49] 此外，公共图书馆还肩负着弥合数字鸿沟的任务：为公众提供获取数字资源的便利，并培养其数字素养。在数字经济中，公共图书馆社会角色的这一转变，使自动化内容分析日益普及所带来的问题更加突出。自动化内容分析软件有望快速洞察大型数字语料库，因此能够为图书馆环境中的研究和批判性思维增添助力。然而，正如本文的案例研究所示，这类工具容易再现上游环节中的偏见，即便其设计初衷是消除偏差。处理数据需要认识到，数据并非中立，而且可能被用于维持不平等的现状。因此，数据既是问题的一部分，也可能成为解决方案的一部分，前提是我们坚持若干原则：其中之一是将数据置于历史与社会背景中考察。图书馆学家可以帮助培养这种意识，并促使用户形成批判且明智的态度。

本研究的一项局限是，由于维基数据、英语版和荷兰语版 DBpedia 的本体和数据质量各不相同，无法对每个数据库使用完全相同的 SPARQL 查询。在某些情况下，我们发现，使用相同的查询文本或类似的属性会偏向某个数据库，却会使另一个数据库返回空结果或有限结果，尽管其中存在可用数据。[50] 由于我们的研究和工具开发聚焦于媒体研究和图书馆情境，因此我们优先获取相关数据，即使这意味着要针对每个数据库使用不同的 SPARQL 查询。

未来的工作中，我们计划进一步分析底层本体中的偏差。这包括：(a) 组合使用多个知识来源（而不是在它们之间做排他性选择），这需要进一步做出设计决策，尤其是在两个来源的信息相互冲突时；以及 (b) 调查维基百科的编辑历史，以“重现过去某个时间点 X 的知识”，补充“当前视角”。我们还计划进一步加深对多样性、偏见与公平性之间关系的理解（包括将其形式化）。

总的来说，我们的研究结果表明，作为工具开发者，我们需要监测知识来源中可能存在的偏差，并研究这些偏差会如何影响“下游任务”的结果（例如多样性的计算与呈现），以及用户感知。除此之外，导致此类过度代表的创作过程本身也是一个非常值得研究的话题。在未来的工作中，我们将研究这两个问题，并将其与我们关于大型语言模型偏见的研究联系起来。[51]

不应忽视的是，自动化工具无法识别和理解语境，这会加剧这些数值考量带来的影响。因此，我们还应将数值结果和分类结果视为进一步开展文本分析的起点，并让用户参与意义建构——无论是普通公民、研究人员、代表特定目标群体的组织，还是希望为新闻报道寻找新角度的记者——都可以通过多样性搜索器的交互界面参与其中。

---

49 同上。[50]例如，我们在英文 DBpedia 和维基数据中使用的查询条件是：实体类型为“政党”，且国家为“比利时”。该查询在荷兰语 DBpedia 中没有返回任何结果，尽管相关实体确实存在。作为替代方案，我们查询了荷兰语 DBpedia，查找那些作为另一实体的政党属性出现、且国家为“比利时”的实体。[51] Delobelle、Tokpo、Calders 等：《使用偏置标尺测量公平性》，2022年

这一理念可以也应该进一步拓展。未来的研究建议采用双管齐下的策略：（a）为特定任务找出并使用最合适的本体；同时（b）向用户清晰说明该本体的属性及其不足之处，使用户了解与该数据集（以及任何数据集）相关的挑战。

致谢

我们感谢 Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO) 资助 DIAMOND 项目，项目代码为 S008817N。

参考文献

关于 DBpedia，DBpedia，（无日期）。<https://www.dbpedia.org/about/>，访问日期：2022年8月28日。

关于 DIAMOND，2018年1月7日， <https://soc.kuleuven.be/ims/diamond/diamond-about>，访问日期：2022年8月27日。

Abián, D.; Guerra, F.; Martínez-Romanos, J. 等：《Wikidata 与 DBpedia：一项比较研究》，载于：Szymański, Julian；Velegrakis, Yannis（编）：《结构化数据源上的语义关键词搜索》，第 10546 卷，Cham，2018年（《计算机科学讲义》），第 142–154 页。在线：<https://doi.org/10.1007/978-3-319-74497-1_14>，访问日期：2022年8月25日。

Amrapali Zaveri、Anisa Rula、Andrea Maurino、Ricardo Pietrobon、Jens Lehmann、Sören Auer：《关联数据质量评估：综述》。《语义网》7(1)：63-93（2016年）

Auer, Sören；Bizer, Christian；Kobilarov, Georgi 等：《DBpedia：开放数据网的核心》，载于：Aberer, Karl；Choi, Key-Sun；Noy, Natasha 等（编）：《语义网》，第 4825 卷，柏林、海德堡，2007年（《计算机科学讲义》），第 722–735 页。在线：<https://doi.org/10.1007/978-3-540-76298-0_52>

Bettina Berendt；Özgür Karadeniz；Stefan Mertens 等：《超越“平等”的公平性：多样性搜索器——检测并提升新闻媒体中社会政治行动者的表征的工具》，载于：《2021年万维网大会》附属论文集，斯洛文尼亚卢布尔雅那，2021年，第202–212页。在线：<https://doi.org/10.1145/3442442.3452303>，检索日期：2022年8月25日。

皮埃尔·布迪厄：《社会空间与群体的生成》。Theory and Society 14（6），1985年，第723–744页。

David Deacon；John Downey；David Smith、James Stanyer 和 Dominic Wring。《2017年全国新闻媒体 对选举的报道》。传播与文化研究中心，拉夫堡大学报告 *4*：*2017*年*5*月*5*日—*6*月*7*日 传播与文化研究中心，拉夫堡大学。在线：<https://blog.lboro.ac.uk/crcc/wp-content/uploads/sites/23/2017/06/media-coverage-of-the-2017-general-election-campaign-report-4.pdf>

Delobelle、Tokpo、Calders 等：《使用偏置标尺测量公平性：预训练语言模型偏见度量指标的比较研究》，载于《2022年计算语言学协会北美分会会议论文集：人类语言技术》，美国西雅图，2022年，第1693–1706页。在线：< https://doi.org/10.18653/v1/2022.naacl-main.122>，检索日期：2022年8月25日。

Färber, Michael、Frederic Bartscherer、Carsten Menne 和 Achim Rettinger。《DBpedia、Freebase、OpenCyc、维基数据和 YAGO 的关联数据质量》。由 Amrapali Zaveri、Dimitris Kontokostas、Sebastian Hellmann、Jürgen Umbrich、Amrapali Zaveri、Dimitris Kontokostas、Sebastian Hellmann 和 Jürgen Umbrich 编。语义网 9，第 1 期（2017年11月30日）：77–129。https://doi.org/10.3233/SW-170275。第46–48页

Hofer, Marvin；Hellmann, Sebastian；Dojchinovski, Milan 等：《DBpedia全新发布周期：提升知识抽取工作流的敏捷性与效率》，载于：Blomqvist, Eva；Groth, Paul；Boer, Victor de 等（编）：《语义系统：知识图谱时代》，第12378卷，Cham，2020年（计算机科学讲义），第1–18页。在线：<https://doi.org/10.1007/978-3-030-59833-4_1>，检索日期：2022年2月21日。

Holze, Julia：DBpedia 2021-12快照发布公告，DBpedia博客，09.02.2022，< https://www.dbpedia.org/blog/snapshot-2021-12-release/>，检索日期：2022年8月28日。

Honnibal, Matthew；Montani, Ines：spaCy 2：借助 Bloom 嵌入、卷积神经网络和增量解析实现自然语言理解，2017年。<https://spacy.io/>，检索日期：检索日期：2022年8月29日。

iCandid，《LIBIS研究基础设施概览》，03.05.2022，< https://bib.kuleuven.be/english/libis/projects#icandid>，检索日期：2022年8月25日。

Kitchin, Rob：《数据革命：大数据、开放数据、数据基础设施及其影响》。伦敦：Sage。2014年。

Lehmann, Jens；Isele, Robert；Jakob, Max 等：《DBpedia——从维基百科抽取的大规模多语言知识库》，载于《语义网期刊》6（2），2015年，第167–195页。

Leorke, Dale；Wyatt, Danielle；McQuire, Scott：“不仅仅是一座图书馆”：智慧城市中的公共图书馆，载于：《城市、文化与社会》第15卷，2018年12月，第37–44页。在线：< https://doi.org/10.1016/j.ccs.2018.05.002>。

Mendes, Pablo N.；Jakob, Max；Andrés García-Silva 等：DBpedia Spotlight：为文档网络揭开面纱，载于：《第七届语义系统国际会议论文集——I-Semantics ’11》，奥地利格拉茨，2011年，第1–8页。在线：<https://doi.org/10.1145/2063518.2063519>，检索日期：2022年8月25日。

Piscopo, Alessandro；Simperl, Elena：《谈论 Wikidata 质量时我们谈论什么：文献综述》，载于：第十五届开放国际研讨会论文集

国际开放协作，瑞典舍夫德，2019年，第 1–11 页。在线：<
https://doi.org/10.1145/3306446.3340822>，访问日期：2022年8月29日。

Heritiana Ranaivoson：使用斯特林模型衡量文化多样性，载于：《统计学新技术与新方法》，2013年，第10页。在线：<
https://ec.europa.eu/eurostat/cros/content/measuring-cultural-diversity-stirling-model-heritiana-ranaivoson en>._

Mark Scott：尽管有人高呼审查，保守派仍主导社交媒体，POLITICO，2020年10月26日， <
https://www.politico.com/news/2020/10/26/censorship-conservatives-social- media-432643>，检索日期：2022年8月28日。

Andy Stirling：分析科学、技术与社会多样性的一般框架，载于：《皇家学会界面期刊》4 (15)，2007年8月22日，第707–719页。在线：<https://doi.org/10.1098/rsif.2007.0213>。

维基百科贡献者：Kamer van volksvertegenwoordigers — Wikipedia, de vrije encyclopedie，2022年。在线：<https://nl.wikipedia.org/w/index.php?title=Kamer_van_volksvertegenwoordigers&oldid=61849646>，访问日期：2022年8月25日。

维基数据仪表板 / 维基数据数据模型陈述，Grafana，（无日期），<
https://grafana.wikimedia.org/d/000000175/wikidata-datamodel-statements?orgId=1&refresh=30m>，访问日期：2022年3月1日。

维基数据：统计，维基数据，（无日期），<
https://www.wikidata.org/wiki/Wikidata:Statistics>，访问日期：2022年3月1日

Denny Vrandečić；Krötzsch, Markus：《维基数据：一个开放协作知识库》。载于：《ACM通讯》57（10），2014年，第78–85页。在线：<http://dx.doi.org/10.1145/2629489>。

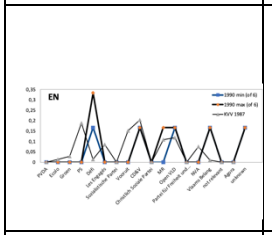
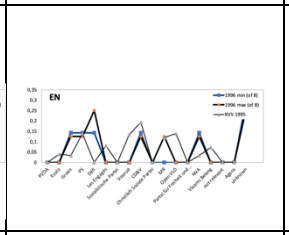
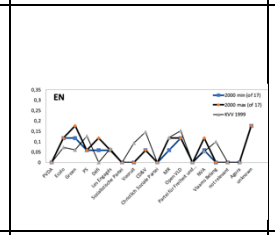
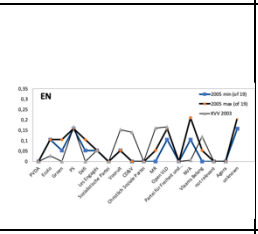
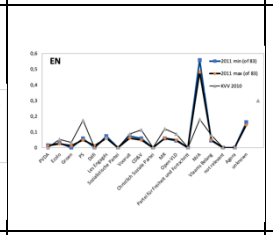
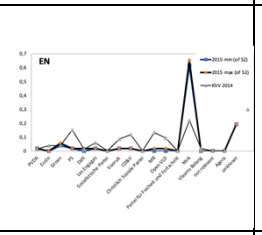
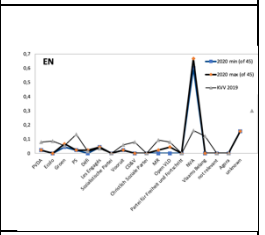
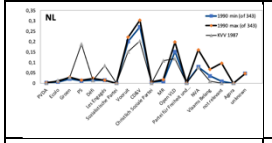
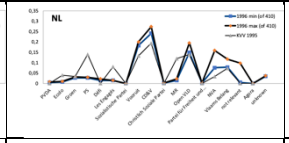
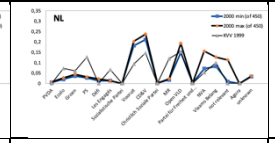
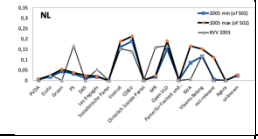
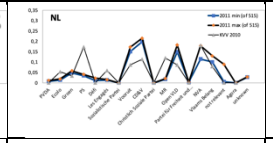
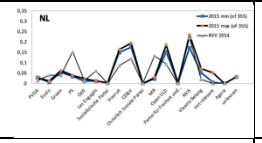
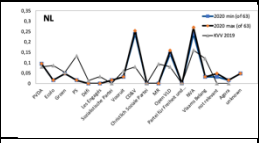
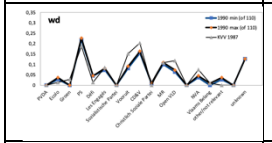
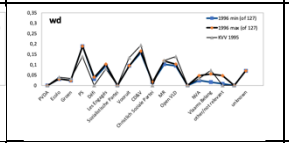
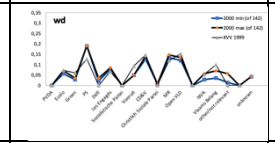
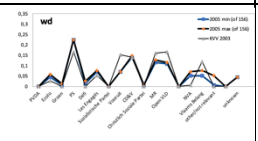
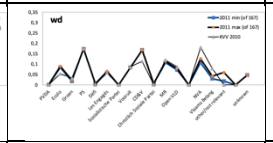
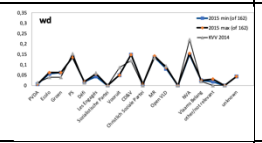
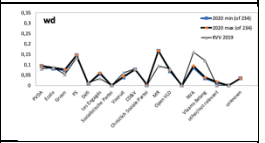
1990	1996	2000	2005	2011	2015	2020
						
6	8	17	19	83	52	45
						
343	410	450	502	515	355	63
						
110	127	142	156	167	162	234

图1. 1990—2020年英语版 DBpedia（上）、荷兰语版 DBpedia（中）和维基数据（下）中各政党的可见度。政党按对齐类别排序（极左、左派、中左、中间、中右、右派、极右、其他、未知），并在各类别内按当前缩写排序。粗线或粗线之间的垂直间隔表示知识来源中的得票份额，细线表示时间上最接近的国家议会得票份额。图表下方的数字表示该知识来源中“年初活跃”的独立行动者数量。。

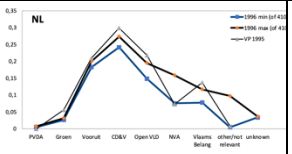
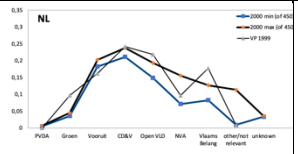
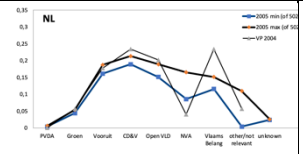
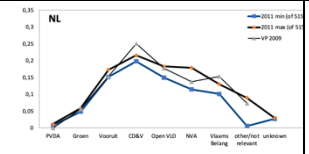
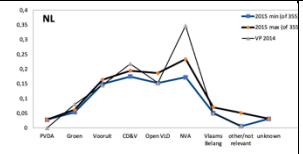
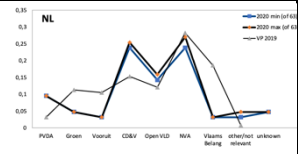
1996	2000	2005	2011	2015	2020
					
410	450	502	515	355	63

图2。1996—2020年荷兰语版 DBpedia 中弗拉芒政党的可见度。政党按对齐类别（极左、左派、中左、中间派、中右、右派、极右、其他、未知）排序；在各类别内，则按当前缩写排序。粗线或粗线之间的垂直空隙表示知识来源中的得票份额，细线表示时间上最接近的佛兰德议会选举得票份额。图表下方的数字表示该知识来源中“年初活跃”的独立行动者数量。