arXiv:2603.11933v1 [cs.DL] 12 Mar 2026

让圣咏计算变得轻松：CantusCorpus v1.0 与 PyCantus 库

预印本

安娜·德沃夏科娃
布拉格查理大学，
捷克共和国

Tim Eipert德国维尔茨堡朱利叶斯-马克西米利安大学

Debra Lacoste
加拿大新斯科舍省哈利法克斯达尔豪斯大学

Jan Hajič jr.*捷克共和国布拉格查理大学
hajicj@ufal.mff.cuni.cz

摘要

数字格里高利圣咏研究数十年来一直得益于一项大型数字资源：Cantus 生态系统。该资源收录了近 90 万首圣咏，涵盖 2000 多个来源。Cantus 数据库的数据模型和 Cantus ID 机制已被另外 18 个圣咏数据库采用，并可通过 Cantus Index 界面统一访问。然而，这些数据此前只能通过各个在线用户界面零散获取，计算方法迄今为止仅有有限机会处理这些海量资源。为突破这一障碍，我们编制了 CantusCorpus v1.0 数据集，整合了截至 2025 年年中以 Cantus Index 为中心的数据库网络中所有可用数据，并提供了用于随数据库扩充而更新该数据集的代码。随后，我们创建了轻量级的 PyCantus 库，以便处理这些数据。PyCantus 将数据模型与 Cantus 代码库解耦，从而能够整合更多圣咏数据源；我们以 Corpus Monodicum 项目的试点数据协调工作为例，展示了这一点。计算圣咏研究前景广阔，而 CantusCorpus v1.0 和 PyCantus 作为基础设施，将使这一领域的研究更加透明、可复现，并让圣咏学者以外的数字人文从业者也能更便捷地参与其中。

关键词 格里高利圣咏 ·计算音乐学 · 数字音乐学 · 数据集协调

1 引言

格里高利圣咏，[1] 即拉丁教会的礼仪单声音乐，是保存最为完好、规模最大的欧洲传统之一，现存手稿超过 30,000 份[Helsen 等人， 2014]而且据推测，中世纪时期可能曾制作过约 500,000 份 [Huglo， 2009]。格里高利圣咏的标准化传统提供了相当易于组织为数据库的材料。[2]在具有读写文化的制度环境中，其神圣性质以及高度规范化地融入礼仪的方式，意味着这一传统大多由界限清晰的复—本单位构成——圣咏（文本与旋律的结合）。其书面记录界定了相关类别，并遵循这些类别：最重要的是体裁（交替圣咏、应答圣咏等），根据其在礼仪中的功能和位置加以界定；礼仪年历中的“圣咏位置”（节日，例如圣诞节或个别圣人的瞻礼，以及该节日中的具体礼仪）；以及高度一致的手稿传统（随后又以相同的书籍编排原则过渡到印刷品）。

这意味着，绝大多数格里高利圣咏的书面记录，都可以明确地被识别为该传统中具体且固定的单元。此外，每个这样的元素都仅凭其文本、体裁和旋律来识别；在大多数情况下，对于给定体裁和礼仪位置中的特定文本，只有一条旋律（这可以从手稿的编排中确定）。因此，人们可以查看来自不同

[1]此外，法兰克-罗马圣咏或罗马-法兰克圣咏源自加洛林时期开始的传统，是法兰克礼仪与罗马礼仪融合的产物。
[2]如需进一步了解格里高利圣咏传统及其结构，请参阅 Hiley [2009]。



图 1：礼仪为一年中每一天的圣咏规定了相应的位置。然而，不同的实践者团体可能会在其中一些位置使用不同的圣咏。因此，每份手稿都记录了特定教会团体在礼仪中安排圣咏的方式。这种安排通过礼仪说明传达，即用红墨水书写的指示，标明瞻礼、日间礼仪或体裁等具体信息。比较多份手稿中处于同一礼仪位置的圣咏，可以发现其中存在差异：主诞前夕的邀请对经（以粉色矩形标出的礼仪说明部分），其圣咏*Hodie scietis* 见于 (a)、(c)、(d)，*Levate capita vestra* 见于 (b)，*Propeest jam dominus* 见于 (e)，而*Christus adveniet nobis* 见于 (f)。（以蓝色标出礼仪说明部分。）

手稿中的任意两条圣咏记录；即使这些手稿相隔数千英里、相差数个世纪，仍然能够明确判断它们是否是同一首圣咏的两条记录。[3]格里高利曲目中的单首圣咏、礼仪，以及记录圣咏在礼仪中使用情况的手稿之间的这种互动，如图 1 所示。

圣咏内部的同质性使其数字化编目相对容易。识别手稿中同一圣咏的不同实例，一直是圣咏研究中的常规工作，尤其在近代史上更是如此，例如索莱姆的编辑工作 [Froger, 1978], Hesbert 于 1963 年编纂的日课对经集 [，以及数字领域中由 Cantus ID 机制实现的工作，尤以 Lacoste [2012]为代表。

使数字领域成为研究圣咏理想场所的第二个因素，是这一传统的规模。这一点在分析方面显而易见：数据量庞大，无法手动处理（通过筛选出一组“具有代表性”的来源，确实可以开展规模更大的研究，例如《Graduale Romanum》版本编纂 [Froger， 1978], 但这仍是一个历时数十年的过程）。编目时也会遇到这一因素：现存手稿的数量和规模意味着这必然是一项长期的协作工作。数据库及其在线访问所带来的规范性，既能确保各项编目工作随着时间推移得以汇集，也能让受时间或资金限制的个人工作在结束后继续为这一领域作出贡献。此外，能够重新查看条目，对于长期保证数据质量至关重要。

早在20世纪80年代，人们就已认识到这一数字化机遇——以及其必要性，并构想了 Cantus 数据库 [Lacoste，2012]（见第3节）。Cantus Index 界面的实现，使人们能够跨所有这些数据库（目前共19个）进行联合检索，彻底改变了圣咏研究 [Lacoste， 2022]。仅在过去一年中，就有61篇文章和8本书报告使用了 Cantus 网站，这足以说明其重要性。[4] Cantus Index 每月活跃用户可能超过2,000人，Cantus 数据库则有8,000人。[5] Cantus Index 联合检索中心目前报告收录了超过1,150,000首编目圣咏。[6]

1.1 我们要填补的空白

尽管网上有大量数字化圣咏记录，计算**圣咏研究**（即前文所说的对数字领域的分析性运用）**却无法充分利用这一资源，**因为数据只能通过各个数据库网站和 Cantus Index 的用户界面逐条访问；而要开展计算研究，就需要一次性加载整个数据库网络的内容。此外，虽然圣咏记录本身相对标准化，来源记录却并非如此：例如，关于单份手稿的来源和年代判定的信息——这些信息保存在来源层级——

[3]但某些较晚出现的体裁除外，例如继抒咏；继抒咏是分节体作品，文本组合方式更为灵活。

[4]据未公开的报告：Jennifer Bain、Andrew Hankinson 和 Debra Lacoste，《年度研究平台与门户进展报告》，加拿大数字研究联盟，2025年11月5日。[5]数据来自 Google Analytics；已扣除估算的机器人活动量。[6] https://cantusindex.org/stats [2. 2. 2026]

对于研究格里高利圣咏曲目的多样性和发展至关重要，但 Cantus Index 网络尚未对这类字段的协议和数据导出格式进行统一编纂和标准化。

近年来，计算圣咏研究领域出现了多项新成果（见第 2.3 小节），有望借助规模大得多的数据，使关于圣咏传播与旋律的长期观点建立在更为坚实的实证基础之上。但要实现这一潜力，**必须首先让圣咏数据可供计算处理**。

1.2 贡献

为弥合这一差距，我们贡献了 **CantusCorpus v1.0数据集**。该数据集包含 888,010 条圣咏记录和 2,278 条来源记录，反映了截至 2025年5月20日 Cantus Index 及其附属数据库网络的状态。其中包含用于从 Cantus 生态系统构建数据集更新版本的提取-转换-加载（ETL）代码。

为使对这些数据的程序化操作变得简单、一致且易于访问，我们提供了 **PyCantus 库**。该库实现了 CantusCorpus v1.0 的数据模型（第 4 节），并提供统一的数据筛选接口，可生成静态制品，以支持实验可复现性，并支持在基于 PyCantus 开发的各类应用之间复用数据子集（第 5 节）。

这些数据[7] 依据 Cantus Index 的许可规定，以知识共享 4.0 许可协议提供；PyCantus 库则为开源软件。[8] 虽然 CantusCorpus v1.0 收集自 Cantus 网络，但我们通过在 Cantus Index 代码库上独立实现 PyCantus，使整合其他圣咏资源更加容易。第 6 节以 Corpus Monodicum 项目的数据为例进行了说明。我们的贡献结构如图 2 所示。

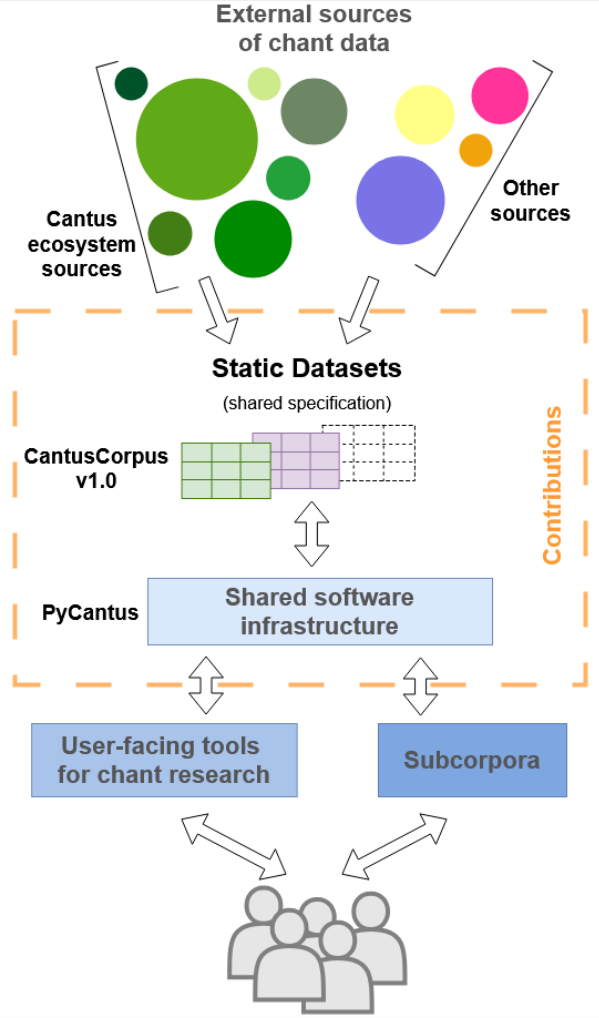

图 2：我们所作贡献的简化示意图。

1.3 为什么选择 CantusCorpus?

仅仅汇编一组现有的相关圣咏数据库，这样的数据集有何价值?

> 1. **重大问题**。 尽管所有数据记录都保存在连接到 Cantus Index 中心的各个站点上，但这些数据并不容易用于计算处理。即使能够直接访问 Cantus Index 后端（显然，实际上不可能授予这种访问权限），数据仍分散在各个站点上。这不必要地限制了计算圣咏研究，例如无法探究“附加段与中世纪政治的关系 [Eipert 和 Moss, 2023a] 是否适用于其他圣咏曲目体裁？”
>
> “圣咏旋律的拼接理论 [Ferretti, 1934年, Levy, 1970年, Treitler, 1975]是否更有可能成立[Lanz 和 Haji˘c jr., 2025]？”或“还有多少种不同的圣咏有待发现？ [Moss等人,2025]”。事实上，在 PyCantus 教程 4 中（见第 5.6 小节），Moss等人 [2025]的一些研究结果已被揭示为相较于 CantusCorpus v1.0 而言，受限数据所造成的假象。若没有一个现成可用的统一数据集囊括所有数据，数据库边界就会阻碍研究——尽管原始数据确实存在。
>
> 2. **易用性**。拥有定义清晰的数据集，也有助于吸引并留住来自音乐信息检索领域或更广泛的数据科学领域、专注于方法研究的合作者。
>
> 3. **透明性**。 Cantus 生态系统持续发展和变化。然而，实证研究需要可复现性。拥有

[7]长期存档版本：http://hdl.handle.net/1 1234/1-6041，开发版本：https://github.com/dact-c hant/CantusCorpus[8] https://github.com/dact-chant/PyCantus

一个通过数字对象标识符或类似机制明确标识的数据集，对于复现实验至关重要。

4. **标准化**。只有在尝试汇集整个 Cantus 网络中的数据时，我们才会发现：尽管数据模型定义明确，各数据库在如何为单条圣咏记录字段赋值，以及公开哪些信息方面，仍存在显著差异。理解这一点的“副作用”之一，是为 Cantus 生态系统进一步实现标准化制定了路线图。CantusCorpus 的抓取基础设施使得创建数据集的新版本变得轻而易举；随着 Cantus 网络逐步采用更受控的词表，新版本还可以扩充数据集，并可能纳入这些词表。

5. **推广**。在圣咏研究之外，Cantus 生态系统也是数字音乐学领域规模最大的资源之一，或许仅次于 RISM。因此，让人们能够轻松着手使用这些数据，在教育环境中也很有价值。随着数字人文（DH）项目日益普及，让圣咏成为兼具人文与技术背景的学生感兴趣的 DH 研究主题，将有助于维系圣咏研究本身。

1.4 为什么选择 PyCantus?

在阐述了将数据库中的各种内容汇编成数据集的价值之后，我们现在说明 PyCantus 库的合理性。

1. **易于使用**。处理 CSV 文件的方式不计其数，而专门的库能够规范处理圣咏资料时的常见步骤，使人们更容易阅读和复现自己的研究成果，也更容易阅读和复现其他圣咏学者的研究成果。

2. **可复现性**。PyCantus 的筛选机制确保能够在更新后的数据集上运行实验，并始终对圣咏和来源采用相同的纳入/排除标准；此外，预处理程序也可以作为静态配置文件共享。这补充了前文提到的透明性目标。

3. **独立性**。我们实现 PyCantus 的数据模型时，并未将其与 Cantus 代码库本身绑定。圣咏已被如此充分地形式化，以至于对这一传统的不同概念化方式实际上可能彼此兼容，因此 PyCantus 数据模型可以有更广泛的用途。为说明这一点，我们利用 Corpus Monodicum 项目的数据构建了一个试验性兼容数据集；该项目完全独立于 Cantus（见第 6 节）。

4. **易于开发**。多个数字圣咏项目都在通过实现近乎等效的圣咏数据模型来“重复造轮子”：包括 Cantus DB 和 Cantus Index 后端、仍在使用 Drupal 内容管理系统的其他数据库，以及 ChantDigger[9],ChantMapper Dvoˇr´akov´a 和 Hajiˇc jr. [2025]and ChantLab Lanz 等人 [2025], or CEAP Hornby 等人 [2022] 等非数据库应用程序。该数据模型和筛选机制可用于任何基于 Python 的应用程序，同时又是一项轻量级依赖项。

与现有界面相比，CantusCorpus v1.0 和 PyCantus 在音乐学研究中的综合价值，在需要同时结合曲目和来源条件进行查询时最为显著。例如，若要比较 12 世纪至 14 世纪的主要圣母节日曲目，并检验熙笃会对这些节日曲目的影响，通过 Cantus Index 用户界面汇编所需的子语料库需要数小时的复制粘贴（还得绕道访问各个数据库网站查询手稿元数据）。之后，还必须针对不同节日的对照组重复这一过程。使用 PyCantus，只需创建包含相应纳入/排除标准的筛选文件和分析脚本，再分享给其他人复现即可。

PyCantus 也足够灵活，可用于圣咏研究以外的场景。计算人文学或文化人类学研究者可能希望比较格里高利圣咏曲目的多样性与其他事物的多样性，例如爱尔兰民间音乐聚会[10] 或现代都市民间故事，[11] 从而更好地理解塑造口头传承的因素。数字人文专业的学生或许能借此认识到，历史数据并不总是符合教科书对某种传统的定义。PyCantus 设有专门面向非圣咏学者的文档章节，将帮助这类用户入门。此外，它还提供了基本圣咏处理任务的参考实现，有助于避免该领域的一些陷阱 [Lanz 和 Hajiˇc jr., 2023]并使这一广阔而迷人的领域能够应用于更广泛的语境。

1.5 文章结构

我们首先回顾数字圣咏研究，以便将这些贡献置于背景之中（第 2 节）。随后，我们介绍 Cantus 数据生态系统，CantusCorpus数据便源自该系统（第 3 节），并介绍 CantusCorpus v1.0数据集本身（第 4 节）。接着，我们将介绍 PyCantus 库（第 5 节），并展示它如何通过数字版项目 Corpus Monodicum 的试点数据集，整合 Cantus 生态系统之外的其他圣咏数据来源（第 6 节）。最后，我们将讨论这项工作的价值与局限（第 7 节）。

[9]https://github.com/timeipert/chantdigger-restored[10]https://thesession.org/[11]https://digifolk.eu/

2 相关工作

CantusCorpus v1.0 和 PyCantus 在数字圣咏研究的更广阔环境中处于怎样的位置?

2.1 数字圣咏数据的来源

CantusCorpus v1.0 和 PyCantus 均衍生自**Cantus“生态系统”**：Cantus 数据库 [Lacoste,2012] 以及另外 18 个数据库组成的网络，这些数据库采用了与 Cantus Index 联合搜索界面相配套的同一数据模型 [Lacoste, 2022]。这是迄今规模最大的数字圣咏数据资源；它每年持续新增数万条圣咏记录。鉴于其核心地位，第 3 节将详细讨论 Cantus 生态系统。

我们具体以 CantusCorpus v0.2 [Cornelissen 等人, 2020a], 为基础，该语料库衍生自 Cantus 数据库。该语料库一直是可复现计算圣咏研究的事实标准，因为迄今为止它是唯一可用的数据集 [Cornelissen 等人，2020b,a, Lanz 和 Hajiˇc jr., 2023, Lanz 等人，2025, Lanz 和 Hajiˇc jr., 2025, Hajiˇc jr. 和 Moss, 2025]。它包含截至 2020 年 Cantus 数据库的全部内容，共计 497,071 条圣咏记录。然而，它没有收集来源信息，也未纳入 Cantus 以外数据库中的圣咏，因此未能充分利用 Cantus 生态系统的广度。

另一个重要的数字圣咏编目项目是**Corpus Monodicum**。[12] 这是一个为期 16 年的圣咏编纂项目，自 2013 年以来一直在开发数据库[13] 和数字工具：Monodi 编辑器 [Eipert 等人， 2019], 与之配套的 MonodiKit 库 [Eipert 和 Moss, 2023b],以及用于音程记谱光学音乐识别的 OMMR4All 工具 [Wick 和 Puppe，2019年，Hartelt 等人， 2024]。迄今为止，该项目已积累了 5,000 多首完整转录的旋律。尽管其规模比 Cantus 数据库小一个数量级，Corpus Monodicum 代表着一项完全独立的数字圣咏研究工作，在目标和数据库结构上均有所不同。因此，它提供了一个机会，用于检验 PyCantus 的设计及其数据模型能否推广到 Cantus 生态系统之外，并为其他拥有圣咏数据的网站提供范例。第 6 节将介绍这场针对 PyCantus 的试点“严峻考验”。

此外，也有针对特定研究问题的小型圣咏数据集，例如 CorpusTroporum 数据集 [Eipert 等人，2025年，Eipert 和 Moss，2023a], 以及小型圣诞节和进堂咏数据集 [Hajiˇc jr. 等人, 2023, 2025]。

[12]https://www.adwmainz.de/forschung/projekte/corpus-monodicum-die-einstimmige-musik-des-lateinischen-mittelalters-gattungen-werkbestaende-kontexte/informationen.html[13]https://corpus-monodicum.de/

2.2 面向当代圣咏歌者的数字资源

此外，还有一些圣咏数据库是为当代歌者和礼仪需求而设计的，而非用于研究。其中最重要的是 **GregoBase**，[14] 它收录的圣咏曲目主要来自索莱姆修道院自 20 世纪初以来出版的现代格里高利圣咏版本，共计 18,930 条圣咏记录。该数据库使用 GABC 编码，可通过 chant21 库导入基于 Volpiano 的 Cantus 生态系统 [Cornelissen 等人，2020a]。GABC 格式也用于通过基于 TeX 的 Gregorio 软件处理流程进行排版，[15] 并用于 Neumz 项目的 scrib.io 编辑器。[16] Neumz 项目包括大规模录制当代圣咏演唱，总时长超过 7,000 小时。这些材料正由 Repertorium 项目进一步处理。[17] 其他应用程序（如 SquareNote）也提供当代礼仪的数字材料，[18] 还有 jgabc 工具包。[19]

这些数字资源的组合能够很好地满足礼仪需求，但与中世纪圣咏研究的相关性不高。

2.3 计算圣咏研究

近来，人们对运用计算方法研究格里高利圣咏表现出浓厚兴趣。以往的人工分析结果必然只涉及有限范围的曲目选集，如今可以通过计算方法在更大范围内重新考察，并使分析更加透明，从而相应地确立更清晰的“证据标准” [McBride, 2025]。

这项研究大多聚焦于旋律。Narmour 提出的旋律拱形假说 [1992] 已在圣咏中得到验证[Cornelissen 等人，2020a]。

先前提出的旋律“方言” [——其中最著名的是“东法兰克”与“西法兰克”方言538], [（索莱姆修士，1960年；Glasenapp，2020年；Hornby, 2002年；Hiley，1993年，第537页；Wagner，1925年；Blachly, 1990],）——已通过系统发育方法重新审视 [Hajiˇc jr. 等人, 2023, 2025]。

圣咏拼接法 [Ferretti，1934年，Levy， 1970], 认为格里高利圣咏旋律由稳定的旋律片段词汇构成；这一观点曾引发批判性争论 [Treitler，1975年，Hiley，1993年，第 74–75], 页。近期，基于对整部手稿的贝叶斯非参数方法实证研究表明，这种观点极不可能成立 [Lanz 和Hajiˇc jr.， 2025] 。然而，具有代表性的旋律模式确实有助于描述曲目结构 [Helsen 等人, 2021], ，也可用于按传统对圣咏旋律进行分类，例如奉献咏 [van Kranenburg 和 Maessen,

[14]https://gregobase.selapa.net[15] https://gregorio-project.github.io/[16] https://scrib.io/[17]https://repertorium.eu/ [18] https://squarenote.co/[19] https://bbloomf.github.io/jgabc/faq.html

2017]。现有工具能够在无谱线记谱法（没有谱线的纽姆谱）中近似发现旋律模式[Hornby 等人， 2022]，，也能够处理处于过渡阶段的阿基坦记谱法 [Phan 等人， 2025]。基于片段并运用机器学习的调式分类 [Cornelissen 等人，2020b，Lanz 和 Hajiˇc jr.，2023，Lanz 和 Hajiˇc jr.， 2025]已被证明优于基于波爱修斯对终止音与音域的定义的音乐理论方法 [Atkinson，2008年，第1]，章，以及基于音高轮廓和音级轮廓的方法[Huron 和 Veltman， 2006]；格里高利圣咏调式与“圣咏传统”记忆之间的联系 [Atkinson，2008年，第3] 章，近期已通过实证研究得到证实[Lanz 和 Hajiˇc jr.， 2025]。

在绘制曲目地图时，为了重新审视将来源归类为传统这一思路， [Ottosen、 2007]，网络模型已被用于 [Eipert 和 Moss，2023a，Dvoˇr´akov´a 和 Hajiˇc jr.， 2025] 并结合地图可视化，以发现不显而易见的曲目传统。

最后，针对圣咏的光学音乐识别（OMR）长期以来受到广泛关注 [Helsen 等人，2014年，Fujinaga，2019年，Vigliensonietal.，2019年，Martinez-Sevilla 等人，2023年，Hartelt 等人，2024年，Fuentes-Mart´ınez 等人，2026]。然而，就我们的工作而言，OMR 位于数据库的“上游”，是实现自动化数据录入的一种途径。

3 Cantus 数据生态系统与 CantusCorpus 政策

Cantus Index 网络中的中世纪圣咏资源，是圣咏研究与探索不可或缺的工具。通过 Cantus Index 中心互联的19个数据库和圣咏研究网站，提升了相关资料的可查找性，并以多种方式汇集数据，从而推动了对圣咏的传统研究。要理解 CantusCorpus 中的数据，需要先简要了解 Cantus 生态系统。

在 Cantus 生态系统数据库中为圣咏编目的过程包括：识别手稿中的各个曲目单元，确定其在礼仪中的功能和位置，转录其文本（也可选择转录旋律）；最重要的是，分配 Cantus ID，以便汇集所有已编目来源中同一圣咏——即格里高利圣咏曲目中的一个单元——的目录记录。此过程如图3所示。

3.1 Cantus 生态系统的历史

Cantus 生态系统的核心是 Cantus 数据库的元数据模式及其原则。该数据库于四十年前创建，当时名为“Cantus：格里高利圣咏数据库”。

这一切始于1984年，Ruth Steiner 在美国天主教大学（位于华盛顿哥伦比亚特区）用个人计算机开发 Cantus 数据库，之后又在大型机上用 COBOL 开发。1997年，Terence Bailey 将其移交给西安大略大学，并将其迁移至 Microsoft Access，从而支持更丰富的注释。

该项目随后进一步发展。2011年，它在 Dru-pal 系统中重建。尤为重要的是，圣咏文本主表与手稿清单记录分离开来，Cantus Index 则围绕该主表建立为一个独立网站，旨在既作为迄今发现的所有圣咏文本和旋律的在线目录，也作为多个圣咏资源的联邦检索平台。Cantus Index 使研究人员能够建立与 Cantus DB 身份独立的数据库，同时仍可通过单一用户界面，依照共享数据模型贡献数据。因此，Cantus 从一个数据库发展成了一个生态系统。重要的是，这一发展显著扩大了有权最终决定哪些来源优先编目的学者群体，从而尽量减少了偏见，例如基于20世纪重构“原始旋律”的努力而隐含地偏好较早来源这一倾向 [索莱姆修士，1960年，Froger， 1978]。2023年，Cantus 数据库迁移至基于 Python 的 Django 内容管理系统。

Cantus 生态系统是数字人文领域规模化成功的一个范例。与此同时，其复杂的历史意味着数据并非完全一致。尽管 Cantus Index 相关数据库的数据库模式大体相同，但在 Cantus Index API[20]所规定的要求范围内，某些字段的填写方式仍有很大的灵活性，这导致其他一些 Cantus Index 数据库采用了各不相同的编目政策。尽管圣咏传播数字分析项目（DACT[21]）最近成立了 Cantus Index主任委员会，旨在为原本缺乏监管的环境建立治理框架，但要实现 Cantus 网络中的数据库标准化仍将是一项长期工作。

3.2 数据库下游政策

我们发现，Cantus 网络中的各个数据库之间存在许多不兼容之处，涉及圣咏和来源。这些问题大多源于缺乏受控词表，而这又为编辑政策各异留下了空间。这种开放性是维系和发展 Cantus 网络的关键，但也使计算工作更加复杂。节日拼写并未标准化（Agathae, infra oct. 与 Agathae,8，或 Adv., hebd. 4., sabbato 与 Sabbato Hebd. 4 Adv.；使用例如莱文斯坦距离的方法也难以识别这些差异）；来源地的描述粒度不一（具体教堂、城市或国家），作品的起源世纪也各不相同（13th century、c.1250、late 13th century、possibly around 1225-50），等等。

其中一些问题正在积极解决。例如，Cantus 数据库中的缩写标识近期已正式与 RISM 政策接轨。Cantus

[20]github.com/dact-chant/cantus-index/blob/main / README.md[21]https://dact-chant.ca/

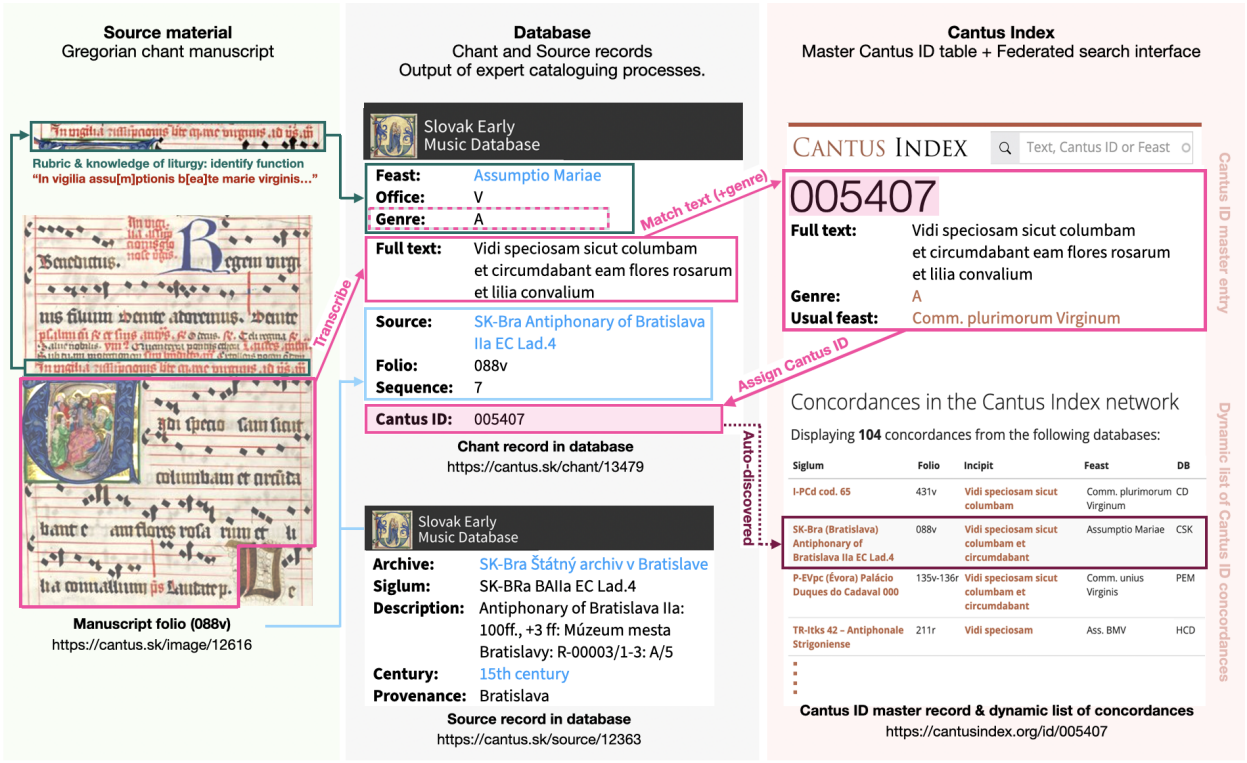

图3：在 Cantus 生态系统中编目圣咏。专家在手稿中识别圣咏（左侧面板），并创建其数据库记录（中间面板上方）。创建圣咏记录的关键步骤是分配 Cantus ID：确定页面上的圣咏曲目单元（**粉色** 过程）。除转录文本外，还需要礼仪专业知识，因为必须正确解读手稿中的缩略注释——礼仪说明——以确定圣咏的礼仪位置与功能（**深绿色** 过程）；结合圣咏文本，这使人们能够在 Cantus Index 的"主记录"中选出正确的 Cantus ID（右侧面板上方）。随后添加指向来源记录（中间面板下方）以及来源内页面（叶码）的链接（**浅蓝色** 过程）。一旦将带有 Cantus ID 的记录添加到 Cantus 生态系统中的数据库，Cantus Index 联合搜索机制（右侧面板下方）就会发现该记录（**深紫色** 过程）。图中提及的各个字段说明见见表1和表2。（来自给定 URL 的截图已为便于阅读而作调整。）

数据库还发布了标注手册，[22] Portuguese Early Music Database (PEM) 也制定了相关指南。[23] 因此，有人可能会认为，像 CantusCorpus 这样的数据集应等到标准化流程完成后再发布。然而，这并不现实。鉴于 Cantus 数据生态系统的复杂性，例如，为节日名称制定受控词表，并在整个网络中实施，可能需要数月时间；此外，古西班牙礼仪或以本土语言演唱的圣咏等更多现象也不断涌现。

因此，与其等待一个虚幻的稳定状态，我们采取在**数据库下游**开展工作的方针。我们的立场是，CantusCorpus 所能持续做到的，不过是反映 Cantus 生态系统内关于圣咏编辑政策的当前共识，并且不应试图取代 Cantus Index 治理内部的编辑流程。相反，我们截取一个快照，记录其局限性，并连同数据集一起提供 ETL 源代码，以便随着数据库的增长及其标准化流程的推进，CantusCorpus 的新版本能够迭代提供。（这也是版本号必须成为数据集名称组成部分的原因。）我们将跨数据库协调中的问题汇总为一份文档（补充文件 S1），随 CantusCorpus v1.0 一同分发。除了描述 v1.0 的局限性之外，它还可以作为进一步标准化的路线图。

[22] https://cantusdatabase.org/documents/[23] https://pemdatabase.eu/node/93860

[24] 参见 https://cantusindex.org/genre 了解体裁缩写。

字段	描述
圣咏链接*	直接指向以下圣咏条目的 URL 链接外部数据库。**唯一标识符**。
开篇词*	圣咏的开篇词。
Cantus ID*	与该圣咏关联的 Cantus ID（例如，007129a）。
mode	圣咏调式 .
手稿代号*	来源手稿的缩写或馆藏（例如 A-ABC Fragm. 1），最好是 RISM。
位置	圣咏在日课中的顺序（第1首、第2首，等等）。
叶码*	圣咏的叶码信息 .
顺序	圣咏在叶码上的顺序。
节日	圣咏使用的节日或礼仪场合 圣咏使用 .
节日代码	用于统一不同拼写节日的附加标识符 其取值为 在 Cantus Index 中有意义。
体裁	圣咏的体裁，例如交替圣咏（A），应答圣咏（R）等。[24]
日课	使用该圣咏的礼仪， 例如晨祷（M）或赞美祷（L）。
srclink*	外部来源的 URL 链接 数据库。
旋律编号	与该旋律关联的编号 圣咏（例如 001216m1）。很少使用。
全文_	圣咏的全文。
旋律	以 Volpiano 编码的旋律。
db*	来源数据库的缩写。
图像	手稿页面图像的 URL 链接 。

表1：圣咏字段概览。星号（*）表示必填字段。

4 CantusCorpus v1.0

CantusCorpus v1.0数据集汇集了 Cantus 风格数据库存在的40年间编目的全部记录，这些记录截至2025年5月20日可通过 Cantus Index 界面获取。该数据集有两个主要组成部分：**圣咏** 和 **来源**。第三个组成部分是抽取-转换-加载（ETL）层：用于收集该数据集的软件。

这些**圣咏**是格里高利圣咏曲目中各个条目的目录记录，例如交替圣咏、应答咏、阿肋路亚咏等，并依照 Cantus Index 标准编目，尤其是 Cantus ID（见图3）。这些记录在圣咏链接中也有唯一标识符，即它们所在数据库中该记录的 URL。

这些**来源**包括包含圣咏的礼仪手稿、早期印刷资料和手稿残片。来源记录包含 CantusCorpus 中基本的音乐史元数据：来源（该书可能的书写和／或使用地点）、可能的书写世纪，以及其*cursus* ——即该书所属的教会机构类型（但并非所有数据库都支持此字段）。圣咏记录相对明确，而来源元数据则可能存在不确定性，具体取决于各个来源的音乐学背景。

4.1 圣咏

圣歌 CSV 文件中各字段的概览见表1。该文件遵循 Cantus Index JSON API。[25]

4.1.1 获取圣咏

CantusCorpus v1.0 中所有圣咏记录均通过抓取 Cantus Index 的 JSON 接口获取，该接口通常用于获取对应关系。对于每种体裁，我们收集其 Cantus ID，并通过端点 URL 依次获取这些 ID。[26] 代码可在数据集代码库中找到。[27]

虽然 Cantus Index 生态系统中列出了 19 个数据库，但通过 JSON API 遍历所有 Cantus ID 后，我们发现圣咏记录仅来自以下 10 个数据库：

- Cantus 数据库[28] （CD）
- Medieval Music Manuscripts Online[29] (MMMO)
- 斯洛伐克早期音乐数据库 [30] (CSK)
- 波希米亚圣咏文献来源[31] (FCB)
- 波兰圣咏[32] (CPL)
- 葡萄牙早期音乐数据库[33] (PEM)
- 西班牙早期音乐手稿数据库[34] (SEMM)
- 匈牙利圣咏数据库[35] (HCD)
- 奥地利修道院的中世纪音乐手稿[36] (A4M)
- Codicologica et Hymnologica Bohemica Liturgica[37] (HYM)

[25] https://github.com/dact-chant/cantus-index/blob/main/README.md
[26] https://cantusindex.org/json-cid/{CantusID}
[27] https://github.com/dact-chant/CantusCorpus/tree/main/cantuscorpus_1.0/抓取
[28] https://cantusdatabase.org/
[29] https://musmed.eu/
[30] https://cantus.sk/
[31] https://cantusbohemiae.cz/
[32] https://cantusplanus.pl/
[33] https://pemdatabase.eu/
[34] https://musicahispanica.eu/
[35] https://hun-chant.eu/
[36] https://austriamanus.org/
[37] https://hymnologica.cz/

4.1.2 数据清洗

将抓取到的 JSON 文件转换为各体裁更精简的 CSV 文件（“Excel 格式”），并将它们合并为一个文件后，我们执行了以下清理步骤:

- 根据找到该圣咏记录所经由的体裁列表，统一体裁名称。[38]
- 删除从不同体裁列表收集到的同一圣咏的重复记录（使用圣咏链接作为圣咏记录的唯一标识符）。
- 排除来自其本身并未公开可见的来源的圣咏（见第 4.2.2 小节）。
- 根据其网页界面，重写匈牙利圣咏数据库中圣咏记录的日课值。

关于这些圣咏数据清理步骤的更多详情，请参阅补充文件 S4。

4.2 来源

来源的情况比圣咏要复杂得多，因为 Cantus Index 不对来源进行汇总，因此不像圣咏那样提供 JSON 接口。

4.2.1 获取来源

每条圣咏记录中可用的信息仅限于来源在其所属数据库网页上的 URL 及其手稿代号。该网页展示了部分来源元数据。我们为所有十个数据库的前端实现了数据抓取程序，这些数据库中的圣咏均可通过 Cantus Index JSON 接口获取。

4.2.2 来源数据字段

格里高利圣咏具有“自我说明”的特性，这主要促成了圣咏数据较为统一的数据模式。然而，Cantus Index 各数据库收集来源信息的方式并不一致。我们调查了它们提供的字段，并在图4中重点展示了可用数据的多样性。

我们选取了所有数据库中都包含的字段（见表2），然后添加了 cursus。我们认为*cursus*很重要，因为它与曲目结构相关，而且通常已知晓（因此通常可以在文本来源描述中找到，即使没有将其标注为字段）。随后，我们根据获得的世纪数值自动推导出世纪数值。

[38]https://cantusindex.org/体裁

字段	说明
标题	手稿名称（可使用手稿代号）
手稿代号*	来源手稿的缩写， 可能为 RISM。
世纪	用于标识该手稿所属世纪的文本 来源 ·
来源	来源的原产地或用途。
srclink*	指向外部 数据库的 URL 链接。唯一标识符。
cursus	世俗或修道院的*cursus* 来源 ·
世纪数值	世纪的整数表示。

表2：来源字段概览。星号（*）表示必填字段。

4.2.3 数据清洗

来源记录被汇编成一个 CSV 文件（“Excel 风格”），并实施了一系列数据整理流程。由于来源的数量比圣咏少几个数量级，因此其中可能涉及人工步骤。

- 解决抓取过程中因重定向而未能传递的 HTTP 与 HTTPS 变更。
- 找出并检查元数据抓取失败的来源——有些来源被手动添加（手稿代号缺失但可恢复），有些则被舍弃，同时舍弃引用这些来源的圣咏（srclink URL 返回 拒绝访问）。
- 根据手稿代号找出重复来源，逐一检查并处理每种情况，例如舍弃其中一个（及其圣咏记录），或将同一手稿的部分内容统一到一个 srclink 下，同时确保圣咏记录的交集只保留一份。
- 根据世纪字符串自动推导世纪数值。

有关来源数据经历的数据清洗步骤的更多详情，请参见补充文件S4。

4.3 CantusCorpus v1.0 统计

表3和表4汇总了圣咏记录与旋律，以及来源记录的数量及其完整性。与圣咏文本编目（其中许多是圣经文本，或是其他已知且在曲目中通行的文本）相比，转录旋律既缓慢又困难，因此仍不普遍。只有少数来源标注了*cursus* 值，这反映了实现这一字段的数据库数量（见图4）。

表5更全面地概述了数据在各来源数据库中的分布。Cantus 数据库是该网络中迄今最古老的部分，（目前）仍是规模最大的（略低于全部记录的一半）。Musica Hispanica（SEMM）

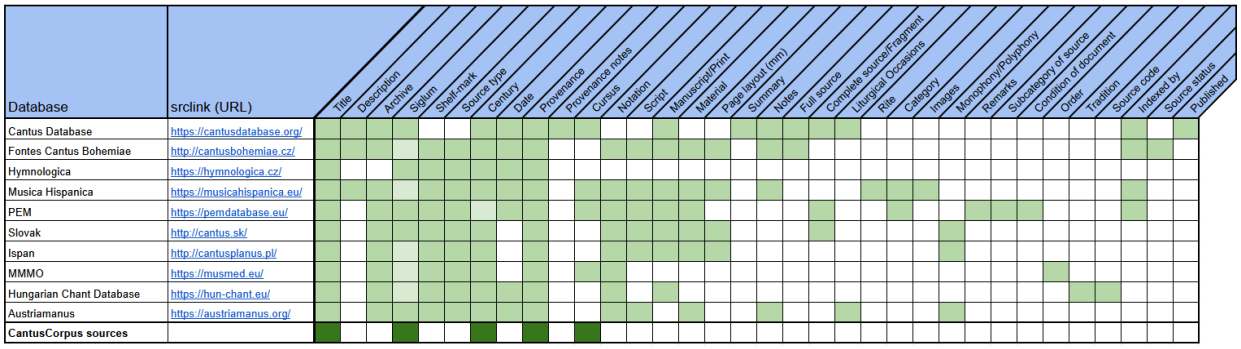

图4：CI 数据库前端对来源元数据支持情况概览。最浅的绿色表示支持以不同名称的字段呈现，最深的颜色表示已选入 CantusCorpus v1.0 的字段。

在其余各项中显得“最为独特”，拥有数量最多的独有 Cantus ID。Codicologica et Hymnologica 数据库（HYM）中独特 Cantus ID 的比例出人意料地高，这源于为以白话（捷克语）书写的圣咏所设的特定语言 Cantus ID。请注意，CSK 和 A4M 包含许多残篇或不完整的手稿。

更多统计数据请参见补充文件 S3。

chants.csv 中的圣咏记录	数量
All	888010
带有 Volpiano 旋律	60588
带有 Volpiano 旋律的 20+个音符	44625

表3：CantusCorpus v1.0数据集中圣咏部分的基本数量值。

sources.csv 中的来源记录	数量
All	2278
全部含 100+首圣咏	508
具有来源值的那些	1606
具有世纪值的那些	2240
具有*cursus* 值的那些	345

表4：CantusCorpus v1.0数据集中来源部分的基本定量值。

5 PyCantus 库

PyCantus 是一个轻量级 Python 库，用于加载和操作 CantusCorpus v1.0 数据集（见第 4 节）。图 2 展示了该库与数据集之间的“分工”。最重要的是，PyCantus 为 CantusCorpus v1.0 实现了一种数据模型。不过，该库是一个独立组件，因此这一数据模型也可用于由其他圣咏数据来源汇编而成的数据集。PyCantus 的功能介绍见教程（见第 5.6 小节）。该库可在 https://github.com/dact-chant/PyCantus/ 获取。

5.1 数据模型

PyCantus 的核心是数据模型（见图 5）。圣咏类和来源类表示数据集中的相应条目，Corpus类对它们进行聚合，而旋律类则支持未来从特定旋律编码中进行抽象。
每个数据类还提供更多便捷方法，例如 Corpus类将自身导出为 CSV 文件，或旋律类清理 Volpiano 表示。

为防止意外修改数据，可在 Corpus 中设置锁，以限制对 Corpus 中圣咏、来源和旋律对象的属性设置器的访问。

5.2 加载器

数据通过 CSV 加载器对象加载到 Corpus类中。该加载器还负责验证。鉴于缺乏受控词表（见第 3.2 小节），它目前只能验证必填字段及其取值是否存在。

5.3 可复现筛选

大多数实验工作流程都涉及对输入圣咏进行筛选：按体裁、节庆集合、来源等筛选。准确描述这些筛选条件对于研究的可复现性至关重要。仅提供筛选后的数据集并非完整的解决方案：后续研究往往还应评估研究发现如何适用于依据相同原则选出的不同曲目集合。

为此，PyCantus 提供了 Filter 类，可将筛选条件导出为 YAML 配置

来源数据库代码	圣咏数	CID 数	唯一 CID 数	# 来源	# 来源（100+）
CD	429982	30350	14662	231	166
MMMO	212231	17479	7503	426	151
CSK	22539	7201	212	542	12
FCB	36103	7889	534	30	29
CPL	30433	7666	143	27	17
PEM	32738	9184	538	305	25
SEMM	104678	23103	11625	487	81
HCD	11278	5374	54	10	9
A4M	2738	2006	12	142	3
HYM	5290	680	323	83	20

表 5：来源数据库中数据分布概览。符号 # 用作“数量”的缩写。数据库代码的缩写见第 4.1.1 小节。标注为 # 来源（100+）的列包含与其关联的圣咏记录超过 100 条的来源数量。

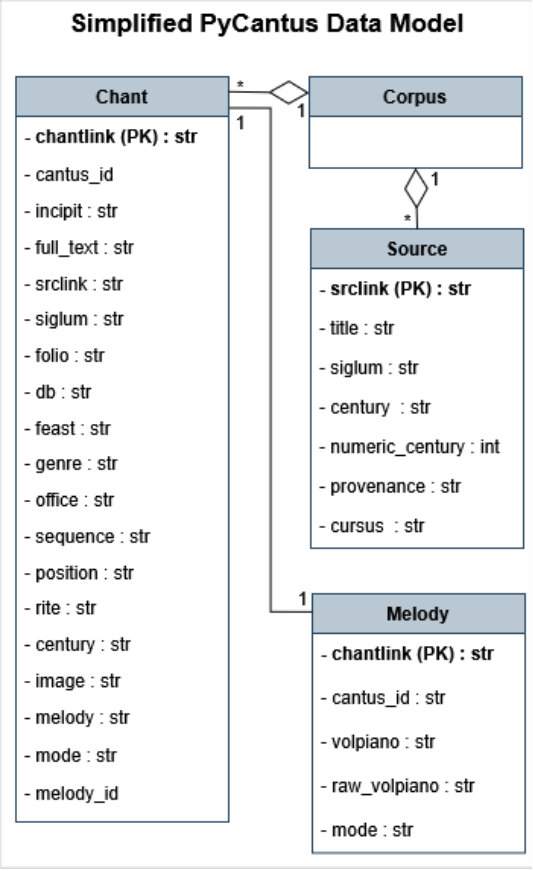

图 5：PyCantus 数据模型的简化示意图（仅显示“内容”属性）。完整的统一建模语言模型见补充文件 S2。

文件（静态且易于分享），并将此类文件应用于语料库。我们还提供了一个网络应用程序，用于创建这些文件。[39][40]

5.4 语料库操作历史

为进一步提高可复现性，PyCantus 会追踪每个 Corpus 类的数据操作历史。该记录可导出为易于阅读的文本，引导用户了解所执行的操作，使其能够轻松地在自己的代码中复现这些操作。

5.5 面向圣咏初学者的文档

可能会对 PyCantus 感兴趣的用户有两类：一类是探索计算方法的圣咏专家，另一类是探索格里高利圣咏的数字人文学者或学生。因此，我们不仅提供技术文档和用户文档，还为格里高利圣咏初学者提供了一份文档。其中包含对相关材料的基本介绍，包括计算方法，从而降低入门门槛。

5.6 教程

最后，我们提供了教程：两篇入门教程（Py-Cantus 基本概念，以及使用圣咏数据的基础知识），以及三篇复现近期在 CantusCorpus v1.0数据集上开展的圣咏计算实验的教程：手稿“社区”检测 [Eipert 和 Moss，2023a；Dvoˇr´akov´a 和 Hajiˇc jr.， 2025]，估计未观测到的圣咏数量 [Hajiˇc jr. 和 Moss，2025年；Moss等人， 2025]，以及按片段将旋律分类到各调式中 [Cornelissen 等人，2020b]。CantusCorpus v1.0的价值已得到

[39]已部署：https://filterforpycantus.owx.cz/筛选器/
[40]Django 模板源代码：https://github.com/DvorakovaA/filterforpycantus

充分证明：在 CantusCorpus v1.0 上复现未观测到的圣咏实验，而不是使用限制更多的 CantusCorpus v0.2，表明[Hajič jr. 和 Moss， 2025] 中的部分讨论是 CantusCorpus v0.2 中弥撒数据有限这一因素造成的假象。

6 超越 Cantus：整合 Corpus Monodicum 数据

PyCantus 旨在使其他来源的圣咏数据能够与源自 Cantus 生态系统的数据实现互操作。为展示这一能力，并为这一过程的实际实施方式提供可操作的见解，我们利用 Corpus Monodicum 的数据构建了一个兼容 PyCantus 的试点数据集。

6.1 Corpus Monodicum

Corpus Monodicum[41] （CM）是一个为圣俗单声部圣咏曲目制作版本的项目。该项目使用 MEI 对圣咏进行编码。作为一个历时 16 年的长期项目，它已经构建了大量自身的计算基础设施：MonodiKit 软件库 [Eipert 和 Moss, 2023b], 、Monodi+ 编辑器 [Eipert 等人， 2019], 以及 OMMR4All 识别软件 [Wick 和 Puppe，2019，Hartelt 等人， 2024]。它拥有完全独立于 Cantus 生态系统的编辑历史、政策和代码库。此外，已编目的圣咏仅有极少部分重叠，而其对德国的关注能够为整个语料库增添大量宝贵材料，其价值类似于 SEMM 对伊比利亚材料的价值。因此，它是检验 PyCantus 数据模型能否作为一般圣咏计算共享基础设施的良好“测试案例”：如果 CM 数据能够以 PyCantus 数据集的形式提供，那么任何圣咏 数据在合理情况下都能做到这一点的可能性就会提高。

主要挑战在于，两个项目中的某些元数据字段并不一致。其中一个字段至关重要：Cantus ID。PyCantus 要求必须提供此标识符，而 Corpus Monodicum 并未记录该字段。

6.2 试点 ETL 流程

CM 数据集可从 Open Science Foundation Repository 获取。[42] 试点子语料库包含 CM 中所有的弥撒专用经对经——这是 CM 对 Cantus Index 中标记为 In、Of 和 Co 的圣咏所采用的术语——但不包括 Graduale Synopticum（GS）子语料库，因为后者基于一个版本，而非中世纪来源。我们选择从这一子语料库着手，是因为它的规范化程度相对较高，且较为熟悉。我们的目标是生成一个 PyCantus 可以顺利导入的 CSV 文件。

我们的 ETL 流程如下：

1. **将 CM 数据集加载到 MonodiKit，并选择子语料库**。 MonodiKit 的默认工作流会返回一个文档对象列表，可直接通过 Python 访问。

2. **查找 Cantus ID**。我们使用 Python 库 rapidfuzz 中的*token sort ratio* 方法，将每份文献与 Cantus Index 中体裁相符的专有对经列表进行模糊字符串匹配（阈值设为 60，该值根据初步测试选定）。常见差异源于规范化方式不同，例如*hierusalem*和*jerusalem*，或是在*dilexistiiustitiam et odisti...*前加入*alleluia*等词。为避免假阳性，只有相似度超过选定阈值时，才会保存匹配项的 Cantus ID。如果返回了多个结果，则会针对节日进行类似的模糊搜索。随后，我们人工确认匹配记录是否正确，并将这些记录及匹配判定写入辅助文件，以供人工核验。需要注意的是，由于我们无法得知 CM 与 Cantus 专有圣咏的确切交集，因此无法评估此方法可能产生多少假阴性。

3.**规范化**。 我们使用 MonodiKit 内置方法将旋律转换为 Volpiano。随后，我们利用人工编制的对应关系，将 CM 来源缩写替换为相应的 RISM 对应项。

4. **导出至 PyCantus 架构**。 我们收集了必填字段，使其与 PyCantus 架构对齐，并将结果写入 CSV 文件。

对于来源，PyCantus 所需的元数据可直接从 CM 获取，因此来源 CSV 文件可以直接导出。所有 ETL 代码均随数据集提供。[43]

CM 数据集包含 498 首专有圣咏（在共计 5474 首圣咏中），其中来自 32 个来源的 420 首圣咏与 Cantus ID 匹配成功，因而被纳入试点的 PyCantus 兼容数据集。其余 78 首圣咏未能找到，可能是由于字符串匹配算法的局限，或是因为这些圣咏可能需要新增 Cantus ID。

6.3 结果

这一工作流程展示了 PyCantus 如何成为连接现有及未来圣咏数据集的枢纽，例如 Corpus Troporum、GregoBase、Graduale Synopticum 或 Comparatio Project。[44] 这些独立项目无法采用相同数据标准，原因各不相同；但对于计算研究而言，它们可以遵循 PyCantus 标准，从而成为 Cantus 之外更大规模圣咏数据生态系统的一部分。

[41]https://corpus-monodicum.de/[42] https://osf.io/mfpkd/

[43]https://github.com/corpus-monodicum/CM-to-PyCantus-Pipeline/tree/master[44]https://comparatio.irht.cnrs.fr/

7 结论与未来工作

CantusCorpus v1.0 数据集和 PyCantus 库“站在巨人的肩膀上”，也就是建立在格里高利圣咏数据的 Cantus 生态系统之上，是迈向大规模数字（及计算）圣咏研究的下一步。这套基础设施将积累了四十年的数据与各个数据库解耦，使人们能够轻松地以编程方式处理这些宝贵数据。因此，我们降低了数字人文研究者的入门门槛，减少了与统计学家及其他技术领域研究者合作时的阻力，并提升了实验性圣咏研究的可复现性。我们尤其认为，让圣咏数据在研究和教学中都易于使用，从而使其从数字人文的角度来看更具吸引力，也是圣咏研究得以持续发展的重要因素。

数据集和库都存在局限性。最突出的问题是缺乏数据验证。这是因为数据库本身许多重要字段（节日、日课、来源等，见第 3.2 小节）缺少受控词表。验证组件需要与 CantusCorpus 后续主要版本的开发同步进行。

要求提供 Cantus ID，会给非 Cantus 数据库带来负担，因为它们需要将数据转换为兼容 PyCantus 的数据集。Cantus ID 模糊匹配器将有助于解决这一问题。我们已针对 Corpus Monodicum 试点数据集成开发了一个概念验证版本（第 6 节）；我们计划进一步完善它，并将其整合到 PyCantus 中。

在 PyCantus 库中，圣咏唯一标识符机制目前遵循 Cantus Index 的惯例，为每首圣咏提供唯一的圣咏链接。随着更多圣咏数据来源接入，这一机制应当更加通用。为避免不同数据库中同一手稿的条目意外重复，还应开发一种类似的“软性”记录匹配器，用于标记可疑的记录对。此外，还应支持非 Volpiano 的旋律编码（主要是 GABC）。

数字音乐学是一种视角，可能有助于回答圣咏研究中的大规模问题，探究广泛传播的中世纪格里高利圣咏传统内部多样性的结构与变化过程。适当的研究基础设施对于这一方法至关重要。我们相信，PyCantus 库和 CantusCorpus v1.0 将成为该基础设施的重要组成部分，使过去四十年收集的大量数字化编目的圣咏能够以全新而令人振奋的方式产生知识。

其他文件

补充文件 S1：

关于规范化问题的报告.*pdf* – 数据集构建过程中遇到的问题概述。HTTP s://github.com/dact-chant/CantusCorpus/blob/main/cantuscorpus_1.0/Report_on_harmoniza tion_issues.pdf

补充文件 S2：

pycantus data model full UML.png – UML 中的完整 PyCantus 数据模型架构。 https://github.com/dact-chant/PyCantus/blob/main/docs/source/_stati c/img/data_model_pycantus_full_uml.png

补充文件 S3：

cantuscorpus statistics.pdf – 关于 CantusCorpusv1.0 的其他图表和计数概览。 https://github.com/dact-chant/CantusCorpus/blob/main/ca ntuscorpus_1.0/dataset_stats.pdf

补充文件S4：

get dataset from scrapes.pdf – 记录从收集的数据准备 CantusCorpus v1.0数据集所采取的全部步骤的文件。 https://github.com/dact-chant/CantusCorpus/blob/main/cantuscorpus_1.0/get_dataset_from_scrapes.pdf

数据可访问性声明

CantusCorpus v1.0数据集可从以下地址获取：http://hdl.handle.net/11234/1-6041。

试用版 Corpus Monodicum 数据集及其配套代码可在此处获取：https://github.com/corpus-monodicum/CM-to-PyCantus-Pipeline/tree/m aster。

源代码

PyCantus 库的源代码可在此处获取：https://github.com/dact-chant/PyCantus。

用于以非编程方式为 PyCantus 创建筛选文件的辅助 Django 模板和简易网络应用程序可在此处获取：https://github.com/DvorakovaA/filterforpycantus/tree/main

利益冲突

作者声明不存在任何利益冲突。

致谢

本研究得到圣咏传播数字分析（DACT）项目的支持，该项目由加拿大社会科学与人文研究理事会提供合作伙伴关系资助（895-2023-1002）。JH 还感谢项目“面向可持续与适应性社会的以人为本人工智能”（注册号：CZ.02.01.01/00/23 025/0008691）的支持，该项目由欧

盟共同资助。部分数据集和计算方法得益于 Corpus Monodicum 项目的持续工作。该项目由美因茨科学院与文学研究院负责，并由德国联邦政府和巴伐利亚州政府资助（https://www.adwmainz.de/）。本文所述研究还使用了 LINDAT/CLARIAH-CZ 研究基础设施提供的计算资源（https://lindat.cz）；该基础设施由捷克共和国教育、青年和体育部资助（项目编号：LM2023062）。
我们还要特别感谢 Jennifer Bain、Anna de Bakker、Ichiro Fujinaga 和 Elsade Luca 的深入讨论。

作者贡献

AD： 概念化、数据整理、形式分析、调查、方法论、软件、验证、可视化、撰写——初稿、撰写——审阅与编辑

TE： 数据整理、形式分析、调查、资源、软件、撰写——初稿、撰写——审阅与编辑

DL： 调查、经费获取、资源、监督、撰写——初稿、撰写——审阅与编辑

JH： 概念构思、经费获取、方法论、项目管理、资源、软件、监督、验证、撰写——初稿、撰写——审阅与编辑

参考文献

Kate Helsen、Jennifer Bain、Ichiro Fujinaga、Andrew Hankinson 和 Debra Lacoste。光学音乐识别与圣咏手稿来源。《早期音乐》，42(4):555–558，2014年10月。ISSN 0306-1078。数字对象标识符：10.1093/em/cau092。URL https://doi.org/10.1093/em/cau092。

Michel Huglo。《记谱礼仪手稿统计调查》。载于 Wolfgang Horn 和 David Hiley 编，《*Antiphonaria: Studien zu Quellen und Ges¨angesMittelalterlichen Offiziums*》，《雷根斯堡音乐史研究》丛书第7卷，2009年。

David Hiley. 格里高利圣咏。剑桥音乐导论丛书。剑桥大学出版社，2009年。

Dom Jacques Froger. 索莱姆修士对《罗马渐进书》的校勘本。《素歌与中世纪音乐学会期刊》，1:81–97，1978年。

Ren´e Jean Hesbert. 日课对经集，第1–6卷，属于《教会文献丛书》大系列：文献来源；*7-12*。Herder，罗马，1963–1979年。

Debra Lacoste. Cantus 数据库：挖掘中世纪圣歌传统。《数字中世纪学家》，7，2012年。

Debra Lacoste。《Cantus 数据库与 Cantus 索引网络》。载于《牛津音乐与语料库研究手册》。牛津大学出版社，2022年。ISBN 9780190945442。数字对象标识符：10.1093/oxfordhb/9780190945442.013.18。URL https://doi.org/10.1093/oxfordhb/9780190945442.013.18。Tim Eipert 和 Fabian C. Moss。《海报：中世纪圣咏集网络中的社群由加洛林政治塑造》。载于《第十届音乐学数字图书馆国际会议论文集》，2023a。

P.M. Ferretti. 格里高利圣咏美学：亦即，格里高利圣咏音乐形式论。格里高利圣咏美学：亦即，格里高利圣咏音乐形式论第sv. 1号。宗座圣乐学院，1934年。URL HTTPS://books.google.cz/books?id=vOWCnQEACAAJ。

Kenneth Levy. 意大利新皈依者圣歌。《美国音乐学会期刊》，23(2):181–227，1970年。

¨Leo Treitler. “拼缀式”圣咏：“拙劣的拼凑”还是“合多为一”？《美国音乐学会期刊》，28(1):1–23，1975年。

Vojtˇech Lanz与Jan Hajiˇc jr. 格里高利旋律、调式与记忆：用贝叶斯非参数方法切分圣咏。载于第*26*届国际音乐信息检索学会会议论文集，第638–646页，2025年9月。doi:10.5281/zenodo.17706544。URL https://zenodo.org/records/17706545。

Fabian C Moss、Jan Hajiˇc jr.、Adrian Nachtwey与Laurent Pugin. 剩下的只有沉默：借助未见物种模型推进计算音乐学。《计算机与人文学科文选》，3:557–574，2025年。doi:10.63744/tP4bLwLkye8B。URL https://doi.org/10.63744/tP4bLwLkye8B。

Anna Dvoˇr´akov´a 和 Jan Hajiˇc jr.《格里高利传统的可视化：ChantMapper》。载于*2025*年音乐编码会议摘要集，第143–151页，2025年6月。数字对象标识符：10.17613/20s0d-gq678。URL： https: //works.hcommons.org/records/g378z-ppk48.

Vojtˇech Lanz、Kristina Szabov´a 和 Jan Hajiˇc jr.《借助 ChantLab 让格里高利旋律的计算研究触手可及》。载于*2025*年音乐编码会议摘要集，第157–163页，2025年6月。数字对象标识符：10.17613/z50gm-qf714。URL：https://works.hc ommons.org/records/z50gm-qf714.

Emma Hornby、Rebecca Maloy 和 Paul Rouse。《圣咏编辑与分析程序：一种分析礼仪圣咏的工具》。《中世纪伊比利亚研究杂志》，14(1):82–95，2022年1月。ISSN 1754-6567。数字对象标识符：10.1080/17546559.2021.2023752。URL：http:// dx.doi.org/10.1080/17546559.2021.2023752.

Vojtˇech Lanz 和 Jan Hajiˇc jr. 文本边界并不能为格里高利对唱咏提供更好的分段。在第十届音乐学数字图书馆国际会议论文集，第72–76页，2023年。

Bas Cornelissen、Willem Zuidema 和 John Ashley Bur- goyne。使用 chant21 研究大型素歌语料库。在第七届音乐学数字图书馆国际会议，第40–44页，2020a。Bas Cornelissen、Willem H Zuidema、John Ashley Bur- goyne 等。素歌中的调式分类与自然单元。在第二十一届国际音乐信息检索学会会议论文集，第869–875页，蒙特利尔，加拿大，2020b。Jan Haji˘c jr. 和 Fabian Moss。知道何时停止：生态学对构建目录、馆藏和语料库的启示。在第十二届音乐学数字图书馆会议论文集，第90–94页，2025年9月。数字对象标识符:10.1145/3748336.3748347。URL https://doi.org/10.1145/3748336.3748347。Tim Eipert、Felix Herrmann、Christoph Wick、Frank Puppe 和 Andreas Haug。支持中世纪单声部音乐数字版的编辑器。在第二届读谱系统国际研讨会论文集，第4–7页，2019年。Tim Eipert 和 Fabian C Moss。Monodikit：中世纪单声部圣咏的数据模型与工具包。在第十届音乐学数字图书馆国际会议论文集，第67–71页，2023b。Christoph Wick 和 Frank Puppe。OMMR4all——用于中世纪音乐记谱法的半自动在线编辑器。在第二届读谱系统国际研讨会，第31–34页，2019年。Alexander Hartelt、Tim Eipert 和 Frank Puppe。中世纪音乐光学识别——历史圣咏的完整处理流程。《应用科学》（*2076-3417*），14（16），2024年。Tim Eipert、Corinna Bongartz 和 Fabian C Moss。《托波鲁姆语料库》数据集：中世纪圣咏附加段元素的数字目录。《开放人文数据期刊》，11(1)，2025年。Jan Haji˘c jr.、Gustavo A. Ballen、Kl´ara Hedvika M¨uhlov´a 和 Hana Vlhov´a-W¨orner。构建格里高利圣咏旋律的系统发育谱系。在第二十四届国际音乐信息检索学会会议论文集，第571–578页。国际音乐信息检索学会，2023年12月。数字对象标识符:10.5281/zenodo.10340442。URL https://doi.org/10.5281/zenodo.10340442。Jan Haji˘c jr.、Vojt˘ech Lanz 和 Gustavo A. Ballen。旋律基因组：将生物信息学用于研究格里高利圣咏的演化。《英国皇家学会哲学汇刊*B*辑：生物科学》，380（1940），2025年12月。数字对象标识符:10.1098/rstb.2024.0274。URL https://doi.org/10.1098/rstb.2024.02 74。John M. McBride。对Buechele等（2023）的评论：跨越鸿沟的沟通——音乐中的物理学有一席之地吗？《实证音乐学评论》，19(2):154–172，2025年3月。ISSN 1559-5749。数字对象标识符:10.18061/emr.v19i2.9783。URL http://dx.doi .org/10.18061/emr.v19i2.9783。Eugene Narmour。《旋律复杂性的分析与认知：期待—实现模型》。芝加哥大学出版社，1992年。索莱姆修士。第四卷：纽姆谱文本，第一卷：手稿分类。载于《罗马渐进咏集：索莱姆修道士编校的批判版》。索莱姆修道院，1960年。John Glasenapp。《不住地祈祷：博普雷反交替圣咏集（巴尔的摩，沃尔特斯艺术博物馆，W. 759–762）中的熙笃会圣咏历时史》。哥伦比亚大学，2020年。Emma Hornby。《格里高利圣咏与古罗马第八调式圣咏：西方圣咏传承个案研究》。阿什盖特出版有限公司，2002年。David Hiley。《西方素歌：手册》。克拉伦登出版社，英国牛津，1993年。Peter Wagner。中世纪早期教会圣咏中的日耳曼与罗曼因素。载于《1925年德国音乐学会在莱比锡举行的第一届音乐学大会报告》，第21–34页，1925年。Alexander Blachly。关于“日耳曼”素歌传统的一些观察。《当代音乐学》，（45-47）：85–117，1990年。Kate Helsen、Mark Daley 和 Jake Schindler。《黏性音型：量化中世纪调式的旋律身份特征》。《实证音乐学评论》，16(2)：312–325，2021年。Peter van Kranenburg 和 Geert Maessen。比较五种中世纪基督教圣咏传统的奉献咏旋律。载于《第18届国际音乐信息检索学会会议论文集》，第204–210页，2017年。URL https://ismir2017.sm cnus.org/wp-content/uploads/2017/10/174_P aper.pdf。Antoine Phan、Martha E. Thomae、Elsa De Luca 和 Francesco Oriol。MEI纽姆谱编码圣咏分析工具。载于《2025年音乐编码会议摘要集》，第36–48页，2025年6月。数字对象标识符:10.17613/jm6rw-btm49。URL https://works.hcommons.org/records/jm6rw-btm49。Charles M Atkinson。《关键纽结：中世纪早期音乐中的音系、调式与记谱法》。牛津大学出版社，2008年。David Huron 和 Joshua Veltman。中世纪调式的认知方法：大小调体系历史先声的证据。《实证音乐学评论》，1：33–55，2006年。数字对象标识符:10.18061/1811/24072。URL https://api.semanticscholar.org/Corp usID:37288543。Knud Ottosen。《拉丁语亡者日课的应答咏与短应答咏》。BoD按需图书有限公司，2007年。ISBN 978-87-7691-186-7。Ichiro Fujinaga。音乐乐谱检索与分析单一界面（SIMSSA）项目：光学音乐

《纽姆记谱法识别工作流》。载于计算机与人文科学研讨会（*Jin-MonCom*）论文集。日本大阪：日本信息处理学会，第*281–286*页，*2019*年。

Gabriel Vigliensoni、Alex Daigle、Eric Liu、Jorge Calvo- Zaragoza、Juliette Regimbal、Minh Anh Nguyen、Noah Baxter、Zo´e McLennan 和 Ichiro Fujinaga。从图像到编码：中世纪与文艺复兴时期音乐的完整光学音乐识别。载于 *2019*年音乐编码会议摘要集，2019年。URL https: //music-encoding.org/conference/2019/abstracts_mec2019/vigliensoni19from%20camera %20ready.pdf。

Juan C Martinez-Sevilla、Antonio Rios-Vila、Francisco J Castellanos 和 Jorge Calvo-Zaragoza。用于音乐与歌词对齐转录的整体方法。载于 国际文档分析与识别会议，第185–201页。施普林格，2023年。

Eliseo Fuentes-Mart´ınez、Antonio R´ıos-Vila、Juan C. Martinez-Sevilla、David Rizo 和 Jorge Calvo-Zaragoza。《对齐的音乐记谱与歌词转录》。模式识别，170:112094，2026年。ISSN 0031-3203。数字对象标识符：https://doi.org/10.1016/j.patcog.2025.112094。URL https://www.sciencedirect.com/science/article/pii/S003132032500754X。