



北京大学

本科生毕业论文

宋代官制结构数字化与

题目：可视化

**Digitalization and
Visualization of Song
Dynasty Government
Structure**

姓名：侯嘉威

学号：1900017878



院系：元培学院

专业：数据科学与大数据技术专业

导师姓名：袁晓如


二〇二四年五月

信息科学技术学院本科生毕业论文答辩评审意见

学生姓名	侯嘉威	本科专业	数据科学与大数据技术	论文成绩 (等级制)	
学生学号	1900017878				
导师姓名	袁晓如	导师单位/ 所在学院	智能学院	导师职称	研究员
论文题目	中文	宋代官制结构数字化与可视化			
	英文	Digitalization and Visualization of Song Dynasty Government Structure			
评审意见	<p>该论文针对现有宋代官职制度研究开展数字化、可视化，是一个典型探索人文与智能结合案例，选题具有重要价值。</p> <p>侯嘉威同学初步探索了宋代官制信息从数字化到可视化的工作流程，以《宋代官制辞典（增补本）》的条目作为数据源进行了自动化探索及分析，并在文本数据结构化开展了工作，初步获得了一定的可视化效果。</p> <p>侯嘉威同学完成整体研究流程，论文工作较为完整。对于相关领域知识具有较好掌握。毕业论文整体较为完整，在具体材料阐述方面还可以进一步优化，整体达到了毕业论文的要求。该论文涉及的方面仍有很大的探索空间，研究潜力较大，希望能继续进行尝试，做出更多贡献。侯嘉威同学答辩中表述清晰，问题回答正确。同意通过答辩。</p>				
答辩时间	2024 年 5 月 16 日 4:45				
答辩地点	理科 1 号楼 1117 室				
答辩评委 签名					

注：至少有 3 个评委签名

北京大学本科毕业论文导师评阅表

学生姓名	侯嘉威	学生学号	1900017878	论文成绩	良
学院(系)	北京大学元培学院			学生所在专业	数据科学与大数据技术
导师姓名	袁晓如	导师单位/ 所在研究所	智能学院	导师职称	研究员
论文题目 (中、英文)	宋代官制结构数字化与可视化 Digitalization and Visualization of Song Dynasty Government Structure				
<p>导师评语</p> <p>(包含对论文的性质、难度、分量、综合训练等是否符合培养目标的目的等评价)</p> <p>该论文针对现有宋代官职制度研究开展数字化、可视化，是一个典型探索人文与智能结合案例，选题具有重要价值。论文初步探索了宋代官制信息从数字化到可视化的工作流程，以《宋代官制辞典（增补本）》的条目作为数据源进行了自动化探索及分析，并在文本数据结构化开展了工作，初步获得了一定的可视化效果。侯嘉威同学完成整体研究流程，论文工作较为完整。对于相关领域知识具有较好掌握。毕业论文整体较为完整，在具体材料阐述方面还可以进一步优化，整体达到了毕业论文的要求。该论文涉及的方面仍有很大的探索空间，研究潜力较大，希望能继续进行尝试，做出更多贡献。</p> <p>导师签名: </p> <p>2024 年 6 月 4 日</p>					

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：

日期： 年 月 日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：

导师签名：

日期： 年 月 日

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以其他方式传播。否则一旦引起有碍作者著作权之问题，将可能承担法律责任。

摘要

本文将视角集中于宋史资料《宋代官制辞典（增补本）》^[1]的官制结构部分，采用图像处理、文本处理和相关的可视化方法，探究材料文本中的条目数据转化方案（如数据结构的设计、古今时间历的映射以及数据提取的方法等），并通过相应的可视化工具或可视化方案实现转化后数据的可视化，以呈现宋朝机构、官职体系的变化，为宋朝政治结构方面的研究和学习以及相关的可视化设计流程提供帮助，为后续更多维度的因素分析奠定基础。

关键词：宋朝，数据机构，元丰改制，光学字符识别（OCR），数字化，可视化

ABSTRACT

This article focuses on the government structure of the Song Dynasty as documented in the *A Dictionary of the Sung Civil Service System*^[1]. Through the application of text analysis and visualization techniques, we endeavor to explore some strategies for transforming the entries into a dataset. Utilizing some visualization tools or methods, our objective is to present the data in a visually accessible manner, highlighting changes in the institution and official systems of Song Dynasty. Our aim is to facilitate the study and comprehension of the government structure of the Song Dynasty and to provide a visualization scheme for the same.

KEY WORDS: Song Dynasty, Data Structure, the Yuanfeng Reform, OCR(Optical Character Recognition), Digitalization, Visualization

目 录

第一章	引言	1
第二章	背景	3
2.1	数字人文	3
2.2	史料数据数字化、可视化相关工作	3
2.3	宋代官制结构可视化需求	5
第三章	任务流程	9
3.1	《宋代官制辞典（增补本）》条目数据描述	9
3.2	图像数据数字化	10
3.3	文本数据结构化	11
3.4	结构化数据可视化	12
第四章	数据分析	13
4.1	数据结构的属性字段	13
4.2	数据处理	15
4.2.1	文字识别	15
4.2.2	数据转化	16
第五章	官制结构数据可视化	19
5.1	数据预处理	19
5.2	树状图实现	20
5.3	机构、官职节点区分	23
5.4	机构、官职节点的折叠和展开	24
第六章	总结和展望	27
6.1	工作内容回顾	27
6.2	不足和提升方向讨论	28
	参考文献	31
	致谢	35

第一章 引言

数字人文领域的兴起，吸引了越来越多人关注古籍可视化的发展。在中国五千年悠久的历史中，经历了众多的朝代的更迭，为数字人文领域提供了丰富的参考素材，涵盖了地理、文学、政治、思想等诸多方面，为可视化工作建立了巨大的文献资料的储备和广阔的研究空间^{[2][3]}。随着可视化研究工作的推进，除了传统前端所需的 HTML、CSS、JavaScript 等技术的支持外，也出现了许多诸如 D3、Echarts、Vue.js 等的可视化开发库和 Web 开发框架，为可视化的实现提供了更高效的技术支持。

在历史材料研究的工作上，传统的路径包括文献研究、田野调查、比较研究以及统计分析方法，过去因为相关辅助工具的缺乏，只能基于单纯的人工手段对相关的文本、图像等数据进行整理，这在整个研究过程中的工作负载是非常巨大的。文本整理、数据转化以及数据校对等工作在一定程度上是可以通过计算机的技术的介入，通过将不同任务切割为独立的工作模块，进行自动化处理的转换，从而最大限度地解放人力资源，使得历史研究工作者可以从大量的重复性工作中脱离出来，将更多的精力投入到研究工作的关键节点中去^[4]。

本次任务主要关注的目标是元丰改制后宋代的中央官制结构名录数据的数字化和可视化，即元丰改制后的中央机构及机构下附属的官职关系的提取和展示。在本次任务中，我们将会把官制信息大致分为三个方面：（1）中央官制；（2）地方官制；（3）品阶勋爵。虽然这三个方面在一定程度上存在着重叠，但在大类上是彼此独立、可以解耦合的。因此，本次工作决定以元丰改制后中央机构的官制结构情况作为切入点进行数字化工作分析和可视化尝试。

古代中国的中央官制体系经历了多个朝代数千年的变迁，在不同时期，其建制和命名都存在着差异，或有所增补、或有所删减、或发生改革、或继续沿用，如秦汉的三公九卿制、唐宋的三省六部制以及明清时期更加彰显中国封建君主集权专制特色的制度，因应不同时期的需求，在每个朝代中也都经历了多次的变更和调整。其中，宋神宗赵顼元丰年间对职官制度进行的变革在历史上就具有重要的影响，该项政治改革在一定程度上改善了宋朝冗官冗员的现象、减少了政府开支、提升了行政效率，从而延续了宋朝国祚。在另一方面，元丰改制更是一个体现古代中国不断推进的削弱相权、加强中央集权的特点的一项改革，其中涉及了大量的机构和官职的增删分合，包含了许多官职由流动、虚领到固定、实掌的官职性质的变更。通过可视化的呈现，为宋史研究者和学习者的探索，带来不一样的视角，以机构官职结构的角度赋予数据的呈现形式以更多的可能性。

本次任务的数据来源于龚延明先生的《宋代官制辞典（增补本）》^①中的各种机构和官职条目。同时，本次工作也得到了历史系和可视化相关领域专家老师同学们的支持和帮助，我们期望能够通过结合自然语言处理、大语言模型工具以及许多相关的可视化实现工具以及领域学者们充分的理论和实践经验，实现对元丰改制后中央官制结构的良好呈现，为后续进一步拓展和完善整个宋朝的官制结构包括中央官制、地方官制、品阶勋爵以及甚至各个朝代的官制结构呈现提供帮助，并作为历史研究工作可以切入的众多维度之一，提供计算机科学领域的解决方案，为史料数据研究在官制结构方向的数字化和可视化工作做出贡献。

本文将会首先介绍可视化工作相关的背景和本次任务的目标和需求；然后提出本次工作的整体框架，并针对此框架，对其中将要面临的问题进行一个初步的梳理和总结；再从数据机构、数字化手段和可视化实现三个方面对当前的工作框架进行展开，分析其中的挑战、相应的处理方法以及所能带来的效果上的优化；最后的篇幅，将会对本次工作进行简单的总结和展望——对本次工作的不足和后续可能的优化方向进行探讨，寻找更多可能带来提升的方向和思路。

^① 为简便起见，下文都将以“辞典”指代《宋代官制辞典（增补本）》。

第二章 背景

2.1 数字人文

“数字人文”(Digital Humanities)这一叫法起源于“人文计算”(Humanities Computing)^{[5][6][7]},最早可以追溯到20世纪40年代末,当时意大利学者罗伯托·布萨(Roberto Busa)与国际商业机器公司(IBM)合作为托马斯·阿奎那(Thomas Aquinas)和相关人员的多达1100多万字的拉丁文作品编制了计算机生成的索引——《托马斯著述索引(*Index Thomisticus*)》,使得计算机技术在语言学领域开始风行,并逐渐开始扩展到人文学科的其他领域中去。随着计算机技术的逐渐发展,其应用对象也逐渐从电子文本逐渐扩展到图像、视频、音频等多媒体。2001年4月一本名为《数字人文指南(*A Companion to Digital Humanities*)》在布莱克维尔出版社出版,此后,“数字人文”这一术语开始得到广泛使用,并迅速取代“人文计算”成为这个计算机学科与人文学科间的跨学科领域的代名词。

数字人文主要的工作方向集中在四个方面^[8]:

1. 建立和维护人文数据库或数据集^{[9][10]},如众多的中西方学术期刊数据库。
2. 根据特定的使用需求将非结构化的数据进行规范化标注形成新的数据集,并开发相关数字人文工具,例如后文将会提到的中国历代人物资料库项目(CBDB, China Biographical Database)^{[11][12]}。
3. 创新数字人文领域的研究方法和研究范式,通过计算机技术的介入,辅助人文学者摆脱繁重的重复性工作,包括数据提取、数据处理、数据生成等等一系列的步骤,从而带来效率上的提升。
4. 通过计算机技术的介入,为人文领域的探索和研究提供新的思路和新方法^{[13][14][15]},在已有的方法论基础上探索数据中更多的特征模式和解读展示的可能性^[16],如虚拟现实技术对于文物资料展示的帮助等,能够从更多维度展示其中的细节和成果。

2.2 史料数据数字化、可视化相关工作

正如大卫·斯特利先生(David J. Staley)在他的文章《历史的可视化》(*Historical Visualizations*)^[17]所说的一样,数据图像不再应该出现在一个研究进程的末尾,视觉通道思维也需要参与到整个设计的构思之中,成为历史研究的另一个新的指导工具。从这篇文章中,作者引用了托马斯·卡莱尔(Thomas Carlyle)对于历史研究因果关系的一

些观点：过去的事件和行为是非常复杂且往往很多时候是同时发生的，从中很难找出一种简单的序列化表达，而用散文 (prose) 表达往往意味着必须限制在某一个维度推进，这是对于历史学者来说导致其研究复杂度的一部分。追求更多维度的表达、支持同时性数据的表达、寻找非线性复杂的网络关系的表达，可视化领域的目标正是在此^{[18][19]}。由此，斯特利先生也表达了对可视化产品成为一项独立研究的最终产品的期待。

不难发现，从很早以前，可视化工作所秉持的逻辑便已经存在，从思维导图，到一些抽象概念的设计，诸如地图、人口分布图等，这些都是众多可视化的形式之一^{[20][21][22]}。通过集合这些设计，提取不同的设计方法对于不同数据的适应性和表达能力，博采众长，赋予整个可视化系统更多维度的表达能力，通过集成，赋予更加强大的表现形式，赋予相关的交叉研究和学科学习更多的可能性^{[23][24]}。

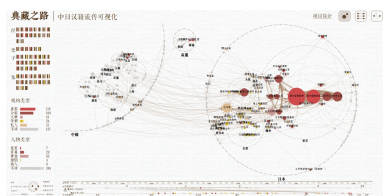
关于中国古籍资料方面的数字人文工作早在 20 世纪 80 年代中期就已经开始了^[25]，最早是针对中国古代文献资料包括《论语》、《全唐诗》等的数据库以及检索系统的建立；后来随着文字识别 (OCR, Optical Character Recognition) 技术的成熟，各类文献的整理、检索方法更是受到了巨大的欢迎^[26]；进一步，随着自然语言处理领域以及语言学领域的发展，伴随着汉语语料库的逐渐扩展，语义分析、词性标注等基于规则以及机器学习的算法模型也开始逐渐出现，进一步为中文古籍资料的分析提供了支持^{[27][28]}。

在中文古籍可视化资料方面，也已经做了不少的尝试^[29]，包括由哈佛大学费正清中心、中央研究院历史语言研究所和北京大学中国古代史研究中心发起的中国历代人物资料库项目 CBDB(如图 2.1)，其中收录了“超过 25000 人之多的传记及谱系资料、4500 条以上的书目资料，以及多种以地理信息为参照系的对象和功能”^①，此外还有一些文献标注^{[30][31]}、古籍流传^[32]和集散路径的可视化等工作 (如图 2.2)。

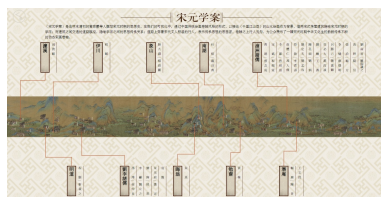


图 2.1 CBDB 在线网页人物传记界面

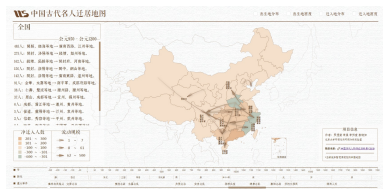
① 引自 CBDB 官网：<https://projects.iq.harvard.edu/chinese/cbdb/%E6%AD%B7%E5%8F%B2>。



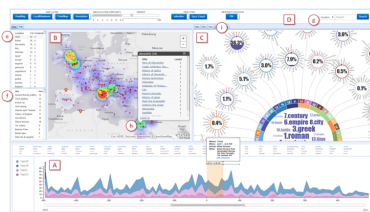
(a) 汉籍流传可视化设计



(b) 宋元学案可视化设计



(c) 中国古代名人迁居地图可视化设计



(d) VAIroma 可视化界面

图 2.2 一些可视化设计样例。(a)-(c) 截取自北京大学可视化与可视分析实验室的系列科研中的相关工作，以及北京大学开设的研究生和本科生可视化课程中部分优秀学生课程设计作品^[33]，(d) 截取自 VAIroma 可视化系统论文^[34]。

2.3 宋代官制结构可视化需求

“两宋 (960-1276) 官制，承隋唐、五代之旧，启明清之新，处于中国官制史上的关键时期”^①，这是龚延明先生在“辞典”总论部分中提到的对于两宋官制重要性的看法，“其繁杂多变，又为历朝之最，给后人了解它的全貌带来很大困难，学者视为畏途”^②。宋代官制的研究，由于别称的误别、概念的混乱以及文本材料庞大带来的校勘上的失误风险的增加，可能会对后来者了解学习宋代官制造成一定的困难，因此龚延明先生提出“本辞典作为抛砖引玉的一种尝试，希望在释疑解惑中，读者能有所取资”^③的想法，为更深入、全面地开展宋代官制研究打开了第一扇门。

“辞典”罗列了约 11600 余条的条目数据，条目内容涵盖了本书大部分篇幅（约 742 页），兼论两宋时期的选官、酬劳、磨勘考课、注授差遣和致仕制度，从源流变迁、行政管理机构概述、官吏管理制度、条目分类、职官术语与典故几个方面对两宋的整体官制结构进行了剖析和释义，成为了对想要了解宋代官制的学者的一本非常有价值的参考书目。

但正如龚先生在“辞典”开头“宋代官制总论”中所说的那样，因为材料的庞大，不同史料记载内容上的不一致以及人工校勘过程中可能存在的失误和混淆的问题，仍然需要相关的历史研究工作者们对本书的内容进行校对和更新，使之成为一本更加权威的两宋官制研究辞典。

① 引自《宋代官制辞典（增补本）》宋代官制总论部分。

② 引自《宋代官制辞典（增补本）》宋代官制总论部分。

③ 引自《宋代官制辞典（增补本）》宋代官制总论部分。

本次任务是基于“辞典”的整理上进行的宋代官制条目的数字化和可视化工程，期望以一种新的形式将“辞典”中整理的名录进行存储及展示，并通过其中的数据转化和生成的过程，对“辞典”中的数据冲突情况，进行一个简单的检查，并通过与相关方向专家的讨论和研究资料的参考^{[35][36]}中，对“辞典”的数据进行初步的清洗^[37]。在本工作之前，在中国古籍文献可视化领域的工作便已经开始了，其覆盖范围包括地理^{[38][39]}、政治^{[40][41]}、人物^[42]、文学等^[43]，有清代职业流动数据的可视化设计、CBDB的历史人物地理分布图、杭州的佛教遗址空间分布可视化及其与社会因素之间作用的分析^[44]等等，所涵盖的数据的维度，所研究问题切入的角度，非常丰富。宋代相关的古籍资料和研究文献资料数量十分巨大，以上的工作也只是众多历史工作中的研究角度的一部分。在此之中，基于朝代的官制结构的可视化数据提取以及可视化尚且未见一个标准的处理方案，在这方面仍然存在着可以探索的空间，作为众多的展示维度之一，官制结构的展示需求尚未得到满足，通过补充此类可视化设计的方案，以结合人物的职官流动状况，继续探索更多的官制变化的线索和依据。

宋朝作为众多朝代中政治经济达到很高的水平的朝代之一，充满着各种各样的政治博弈，不同的派系之间此消彼长，争斗不断，无论机构、官职的增废，还是人员的变迁都非常剧烈。通过“辞典”的条目数据的提炼，将之与人物官职流动数据组合，将这些变化通过可视化的形式呈现，探寻潜藏在其中的关联性，将会是一件非常有意思的事情。

“辞典”中的条目包括机构、官职和物品三个大类，本文将重点集中于机构和官职的部分，根据机构-机构、官职-机构之间的隶属关系构造层次型数据结构的表示，以便对宋代元丰改制后的中央官制结构情况有一个基本的认识。根据前文所列举的“数字人文”部分介绍的主要的工作方向，本次工作的工作方向主要将集中在：

1. 建立合适的数据结构：元丰改制后的中央机构和官职名目数据集的建立。
2. 自动化手段的探索：通过文本处理的技术以及相关的计算机技术手段实现人工处理向自动化处理的过渡，简化部分人工提取过程，探索能够通过自动化技术代替人工重复性工作的手段，为后续的数据处理提供便利。
3. 结构化数据的生成：将“辞典”中非结构化的各名目数据转换成结构化的数据表达，最终形成可用的 Excel 或 JSON 格式。
4. 可视化实现的探索：根据现有的一些可视化开发工具，形成对应时间节点的数据展示，在此基础上，尝试探索不同的潜在的可视化需求，形成一个初步树状层次结构，并进行一些简单的交互性尝试。
5. 数据校对：正确的可视化呈现需要基于正确的数据关系的映射，在一定程度上，构造“辞典”条目的层次数据的过程中，也是一次进行数据检查的过程，通过

树状层次结构数据生成过程中自然产生的数据检查过程，对其中的缺漏和冲突数据进行更新、优化。

第三章 任务流程

本章节将围绕源数据的特征描述、图像数据数字化、文本数据结构化、结构化数据可视化四个方面进行说明。

图 3.1为任务流程概览。

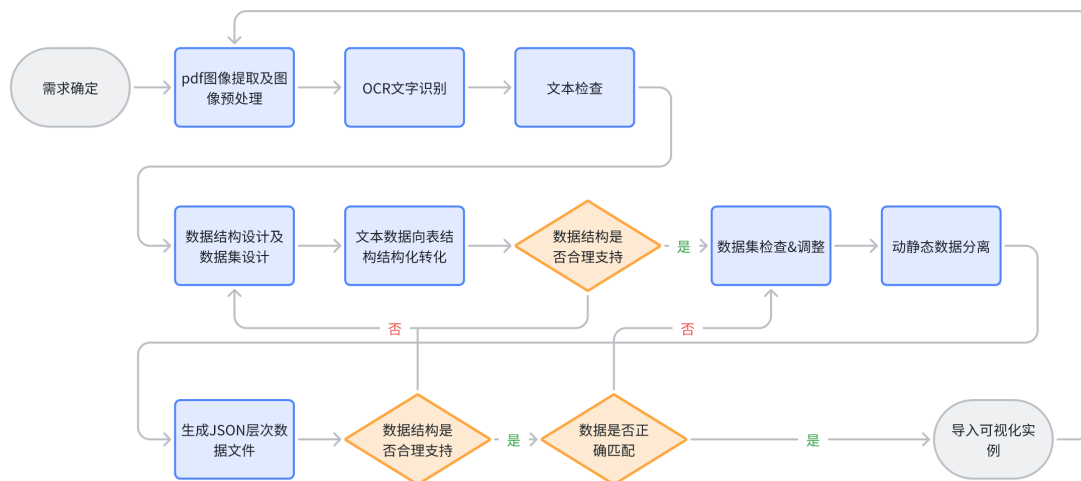


图 3.1 任务流程图

3.1 《宋代官制辞典（增补本）》条目数据描述

本次数字化及可视化任务的数据来源于“辞典”中元丰改制后中央机构名录部分（图 3.2），其范围为直属于皇帝下的主要机构，包括但不限于三省六部制中的“三省”、“六部”、“九寺”、“五监”等大类下的条目。字段上，“辞典”中也已经预设了部分含义项，在条目中的出现顺序为（1）官司或官吏名；（2）职源与沿革；（3）职掌（或职能）；（4）官品（或品位）；（5）编制；（6）简称与别名。^①这六类为“辞典”预定义字段。但是这六类字段在“辞典”中可能存在不同的表达形式，如“简称”、“别称”、“通称”等应属于“简称与别名”字段类中，又如“职源、编制与职能”这一直接包含多个预定义字段情况的字段标题的情况。针对此类形式比较自由的标题命名风格，在本次工作中将会基于预定义的六种字段，对其中的标题字段进行统计并完成标题字段到预定义字段的映射模式，从而加快数据处理的速度。

此外，由于网上没有找到“辞典”的原始电子文本文件，只找到了“辞典”的影印本文件（即图像文件），故直接将此影印本文件作为本次任务的数据源，对于该文件的

^① 同《宋代官制辞典（增补本）》凡例部分的描述。



图 3.2 《宋代官制辞典（增补本）》图像示例

处理也可以作为图像处理工具对文本图像的文字识别能力的简单效果测试，下文将会继续说明对该数据源的处理方案。

3.2 图像数据数字化

图像的提取方面，本次工作使用 Python 的 pdf 处理库 fitz 库的 PyMuPDF 模块对“辞典”中每页的图像进行提取，其中包含控制提取图像分辨率的函数，经过测试，提高图像的分辨率后，文本识别工具识别结果的准确率显著提高。

目前市面上已经存在着一批功能比较完善的文字识别（OCR）工具可以进行图像中的文字的提取，该类工具运用计算机图像处理的方法，通过对图像进行预处理（图像光影的处理、图像倾斜的处理、图像扭曲的处理）尽可能地形成规范化的黑白图像，再根据行列间距状况对字符进行切割，接着利用神经网络模型对切割后的字符图像进行文字识别，最后通过坐标映射来还原文本的排列，并使用诸如规则匹配和机器学习等方法根据预定义的规则或者上下文信息进行文本纠错，通过喂入大量的语料数据来提高 OCR 工具自动识别并纠正错误后生成的文本的准确率。

效果上，OCR 工具和传统的人工提取相比存在着一定的精确度误差，生僻字的识别误差、形近字的识别误差，缺漏字符的情况也依然存在一定的比例，但是通过 OCR 工具的辅助，使得图像数据向文本数据转化的这一过程的工作量被大大减小了，极大

地降低了人工提取的成本。由于该工具对于图像文件大小的限制，前述图像提取的函数中控制图像分辨率的缩放系数最终设置为 3。

调用 OCR 工具将图像数据转换为文本数据后，需要设计符合需求的数据结构将非结构化的文本数据转换成结构化的表格数据，以方便进行后续的可视化层次数据的生成。在这一部分，将会使用一些文本处理的手段，将对应属性所包含的正文文本进行分类映射——将可以明确规则化的属性数据通过正则匹配的手段进行提取。而需要结合上下文参照理解的数据，将进行大语言模型的推理和提取的尝试，目前因为效果并不理想而决定暂时采用人工处理的方案（优化空间很大，需要进行试验）。

由于不同的条目在归纳的粒度、表述的完整性以及依赖关系的分配上是存在一定比例的不一致和留白的，或许对于历史系专家来说这部分属于内化的知识，但是对于并不精于相关朝代官制情况的数据工作者而言，在数据提取和转化工作上，这将会成为一个很大的障碍，将不得不额外引入其他的工具或人力来保证数据的准确性和完整性。所以在上述的两个阶段中，我们的工作仍然需要反复穿插人工校对的部分，以及需要对问题数据进行讨论，参考领域专家的意见进行最终数据细节上的调整。

本次任务中将时间节点定位了元丰改制后，由于“辞典”在数据搜集的过程中参考的文献非常丰富，不同文献的作者的考证结果、写作风格以及文献本身因为保存问题而出现的损耗等等，导致同一个条目下不同的事件发生的时间节点可能存在着一定的模糊和冲突。如果将机构、官职条目的增删、修改操作作为一个事件的话，不同事件记录的时间粒度可能是不一致的，有的可能精确到了日，有的可能仅仅提到了某个时期，由于这些粒度的不一致，也会导致时间映射和实际映射之间的数据冲突。因此，在这部分将会将事件记录的时间粒度统一为精确到公元年，若是该事件只提到某段时期而不是某个精确的日期，参考“辞典”中的处理逻辑，便以该时期的所能映射到的最早精确时间作为映射，并统一到公元年的粒度上。之所以选择公元年作为最终统一的时间粒度，是因为经过不同粒度的尝试后，结果显示精确到年能最大限度地减少机构-机构、官职-机构间依赖产生的时间冲突。本次工作的主要目标在于呈现宋代中央机构的整体官制结构，故选择该粒度作为映射方案，牺牲一定的精度，减少数据冲突。后续若需进行精度的提升，则需要更多的历史系专家参与，通过修改映射的精度，也可以通过其中的数据冲突情况，探寻不同机构、官职信息之间增废时间的关联，从而实现对应历史数据的时间维度上的优化。

3.3 文本数据结构化

关于结构化数据，首先需要包含“辞典”中的六类预定义字段：（1）官司或官吏名；（2）职源与沿革；（3）职掌（或职能）；（4）官品（或品位）；（5）编制；（6）简称

与别名。此外，由于分类的需要，需要对机构和官职的类别进行区分。为了满足溯源和检索的需要，亦需要新增相关的字段保留“辞典”中的条目所在位置的信息以及条目中的引用情况等。

从数据结构设计的角度考虑，一定的字段冗余是有必要的，前述的数据可以保留其“辞典”中的原文正文，另外再添加字段保留其简化后的信息，为后续预览信息和详细信息的可视化需求提供支持。

3.4 结构化数据可视化

当前已有不少的可视化开发工具，包括 D3、ECharts、Vega 等基于 JavaScript 的可视化开发库，给予了可视化设计巨大的支持，很大程度地降低了可视化实现的设计成本。

除此之外，在实现方面，也已经有了不少工作成果——因应不同的可视化需求，提出了相应的解决方案^{[45][46]}。本次可视化任务的目标是进行宋代官制结构的展示，初步可以理解为层次化数据的展示，相关的实现形式有很多，如树状图、树图、冰柱图、旭日图等（如图 3.3），为可视化效果的直观，也因为 D3 开发社区已有的实例的支持，本次任务的可视化选用了树状图的实现。

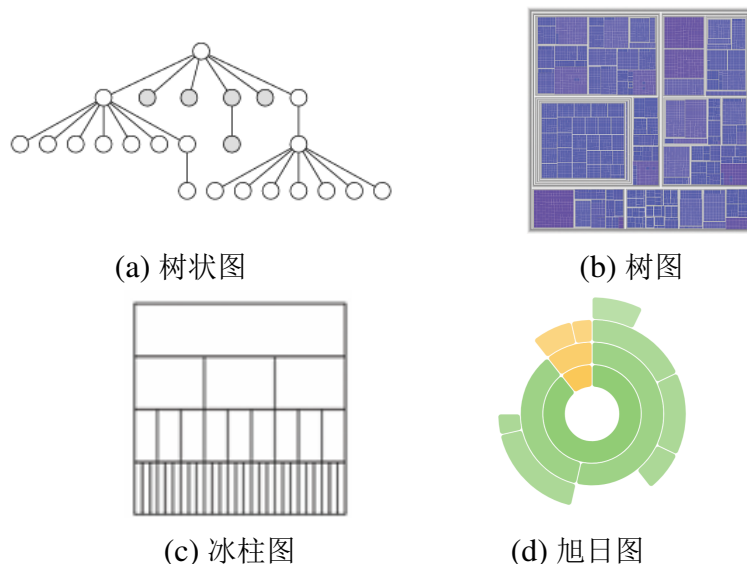


图 3.3 部分层次数据可视化形式样例。(a)-(c) 截取自 *Visualization Analysis & Design*^[47] 第九章的图片，(d) 为 Echarts 开发库旭日图实例。

第四章 数据分析

本章节主要关注数据结构设计和数据提取的部分。

4.1 数据结构的属性字段

数据结构的属性字段的初步设计在上一章节已经概述，在这一部分，将会继续进行详细的说明。根据“辞典”中的属性设置，可以初步拆解出“条目名”、“职源与沿革”、“官品”、“职掌”、“编制”、“简称与别名”这几个字段，都以原文的形式进行直接提取，考虑到后续信息预览的需求，除了“条目名”字段，都会新增一个派生字段将其核心信息提炼出来（如：“简称与别名”下的数据中除了对应的简称与别名外，还会列出出处，在派生字段中会将出处及原文索引等额外内容删减，只保留对应的名称）。

其次，经过讨论，因为溯源的需求，除了保留相关字段的原文文本数据外，新增了“参考文献页数”以及“出处”两个额外字段，以满足检查和检索的需要。

第三，通过对“辞典”的简单了解以及同历史系老师们的讨论中发现，不同的机构、官职本身存在着一些特征，如机构上，可能会有仓库、官署、衙署（实际办公场所，经过讨论后选择作为一个属性独立出来）等区分（衙署独立出来后，其他的机构从特征上基本是一致的，经讨论后暂时不做区分，仅标注为“机构”以同“官职”类区分开来）；官职上，存在着阶官（标定品阶勋爵，确定食邑俸禄所用，无特殊职能）、职事官（有特定工作内容的固定官职）、差遣官（有特定工作内容的派遣官员）、胥吏（机构下实际执行任务的官员）等^[48]，故加入了“类别名”这一分类型属性进行表示。机构和官职事件中，除了增加和删除，还包括更名，故加入了“前身”字段。

此外，还有时间属性，在该属性下细分为了“时间”、“年号纪年”和“公元几年”三种属性，“时间”字段保留原文中的时间表述，后两者根据上一章“数据预处理”部分的原则，保留对应的年号纪年和公元纪年信息，作为每个事件的时间线索——以这个结构为基础，根据不同条目内容中提到的增删情况，将相应的时间属性投射到“始置时间”和“罢置时间”这两个复合属性中去。

最后，根据机构-机构、官职-机构间的隶属、依赖关系，新增“上级机构”（机构-机构）/“隶属机构”（官职-机构）字段保留不同条目之间的依赖线索。因为官职信息中存在着少部分的位次高低之分的表述，故新增了“下级官员”以及“平级官员”字段。

在不同字段的数据结构设计上，“公元纪年”为时间数据，“隶属机构”、“上级机构”、“上级官员”、“平级官员”皆为文本列表，剩余的属性将直接保存为文本数据。

以上就是相关的数据结构的字段设计，后续随着工作进度的推进，也会因应需求的改变继续进行调整。

表 4.1 机构结构字段示例

字段名	含义
条目名	条目名称
类别	条目类别（机构/官职）
衙署	机构下办公场所
职掌文本	原文职掌文本
职掌	对应时期的职掌信息
编制文本	原文编制文本
编制	对应时期的编制信息
编制人数	下辖的官、吏人数情况
简称与别名	原文简称与别名文本
职源与沿革文本	原文职源与沿革文本
参考文档页数	“辞典”中对应页码
出处	条目中引用的文献
下级机构	子机构列表
上级机构	父机构列表
开始时间-年号纪年	该条目存续期开始时间-年号纪年
开始时间-公元纪年	该条目存续期开始时间-公元纪年
结束时间-年号纪年	该条目存续期结束时间-年号纪年
结束时间-公元纪年	该条目存续期结束时间-公元纪年

表 4.2 官职结构字段示例

字段名	含义
条目名	条目名称
类别	条目类别（机构/官职）
职掌文本	原文职掌文本
职掌	对应时期的职掌信息
编制文本	原文编制文本
编制	对应时期的编制信息

续表 4.2 官职结构字段示例

字段名	含义
编制人数	该官职的官额/吏额
官品文本	原文官品文本
官品	官品文本中的官品信息
简称与别名	原文简称与别名文本
职源与沿革文本	原文职源与沿革文本
参考文档页数	“辞典”中对应页码
出处	条目中引用的文献
隶属机构	父机构列表
开始时间-年号纪年	该条目存续期开始时间-年号纪年
开始时间-公元纪年	该条目存续期开始时间-公元纪年
结束时间-年号纪年	该条目存续期结束时间-年号纪年
结束时间-公元纪年	该条目存续期结束时间-公元纪年

4.2 数据处理

4.2.1 文字识别

“辞典”的一页双列格式，在文字密度过大时，OCR 的结果会将本来分属一页两列的文本识别为一页一列的文本，导致其识别结果的顺序出现问题，若是将其交由文本人工检查阶段再处理，将会变成一项非常耗时的工作。因此，在文字识别任务中，第一个需要进行的工作就是将图像按列切割，避免 OCR 工具在双列由于文字密度较大而过于接近产生出现误切分的情况。但是因为数据源文件在扫描过程中倾斜和偏移的问题，暂时没有找到一个统一的标准能够保证所有图片的切割结果都是正确的，所以在此采用自动化处理后进行人工检查重新切割的方案，先对每页的图像进行批量对图片从中间对半切分的处理，再交由人工检查图片是否正确将两列切分开，重新切分质量不好的图片，保证所有切分后的图片都正好分为干净的左列条目及右列条目。

除了使用 OCR 工具之外，本次工作也尝试了使用大语言模型直接对文字识别后的文本进行错误检查，但是当前效果并没有带来很明显的提升，前述的问题仍然会导致大语言模型的推理偏离实际的数据提取需求。从某种意义上来说，大语言模型是否能很好地解决该类问题亦是取决于其调用的 OCR 工具，OCR 的后处理部分已经包含了基于规则匹配或机器学习的算法根据上下文线索进行文本纠错的流程。在这方面，大语言模型在没有喂入合适语料的状态下，并不能显著改善前述 OCR 识别存在的问题。

因此，图像切割的工作最终仍然选择了“自动化切图 + 人工检查重新切分问题图片”这一方案。

在 OCR 工具完成了初步的图像数据转化后，就需要人工介入进行文本数据的检查。由于 OCR 本身存在着一定的误差，部分关键的词句会出现识别错误的情况，如“差遣”识别为“差遣”，“入内内侍”、“侍郎”识别为“人内内侍”、“侍郎”等情况，在提高图片解析度后其结果有显著的改善，但是因为本次任务选用的 OCR 工具在文件大小的支持上有限制，不能无限制地通过提高图片解析度来彻底解决问题，最后的识别结果仍需进行人工检查修改。调用大语言模型进行推理修改，但提升效果并不明显，关键文本信息的识别错误仍然无法纠正，仅仅改善了一些格式和符号上的问题。大语言模型介入对于效果的提升并不大，且本次工作中所使用的 OCR 工具生成的文本并没有根据其源数据字体大小进行标题字段的区分，所以暂时无法实现完全自动化，但仍然保留将大语言模型作为其中的辅助工具的选项，需要继续对其工作流程的控制进行优化。

为了后续检索、修改的方便，需要将条目和标题字段转化为高亮形式标题格式。从体验上，这能极大地改善阅读者的对于文本信息的定位速度和阅读体验^[49]。基于“辞典”中属性字段的命名形式变化不会太复杂的这一假设出发，首先在一个选定门类下统计属性字段的名称集合，再通过正则匹配的方法将对应的属性字段在 markdown 文件中转化为格式化的标题，接着由人工介入进行条目的标题化，因为源数据中属性字段的命名方式比较随意，如“官品”可能表述为“品阶”、“品秩”等，因此也需要随着人工检查的进行，补充缺少的标题字段到预定义字段的映射，逐步提高批量处理的效率。在标题化任务完成后，最后再进行文本内容的检查。以上工作，除了提供了后续表格转化的方便外，标题化的处理对于定位条目和浏览文本信息的体验感相对于直接阅读生文本有了显著的提升，减轻了文本内容修改过程中的认知负担。

4.2.2 数据转化

前文已对数据结构的设计做了基本的描述，所需的字段可以大致分为四类：

1. 直接根据“辞典”预定义字段按标题字段提取的原文数据，如“官品_原文”、“编制_原文”。
2. 需要在从原文数据中提取的数据，如“类别名”、“官品”、“编制”、“简称与别名”这类简化数据，除了简单的文本处理工作，暂时仍需要人工的参与，后续或可进行一些语料上的收集通过大语言模型的辅助生成本部分的表格内容。
3. 需要引入外部知识进行转换的数据——时间属性数据，需要建立一个时间数据到年号纪年表达上的映射，再接着从年号纪年数据映射出公元纪年的数据。

4. 额外的非条目内数据，如“参考文档页数”。

基于以上分类，相关的文本数据结构化转化方案如下：

1. 根据“辞典”中的标题字段直接匹配原文数据。
2. 人工提炼简化数据。
3. 统计在上一步提炼了的文中的时间表达，根据 3.2 的映射逻辑生成对应的年号纪年和公元纪年的映射。
4. 自动化提取过程中根据文档页数标记生成“参考文档页数”字段的数据。

进行该部分的转化后，将会得到“机构条目表”、“官职条目表”、“时间映射表”三个原始表格。由于机构、官职条目存在多个存续期的特性，如“编制”、“职掌”、“官品”等字段上，部分条目可能会在不同存续期时包含不同的内容，所以对应字段的简化字段数据需要因应不同存续期的内容进行保留。所以，根据其字段是否需要发生变化而将字段分为动态字段和静态字段分别存储，由事件来表达不同存续期的变化，这部分将在下文展开描述。

数据处理的总体流程如图 4.1。

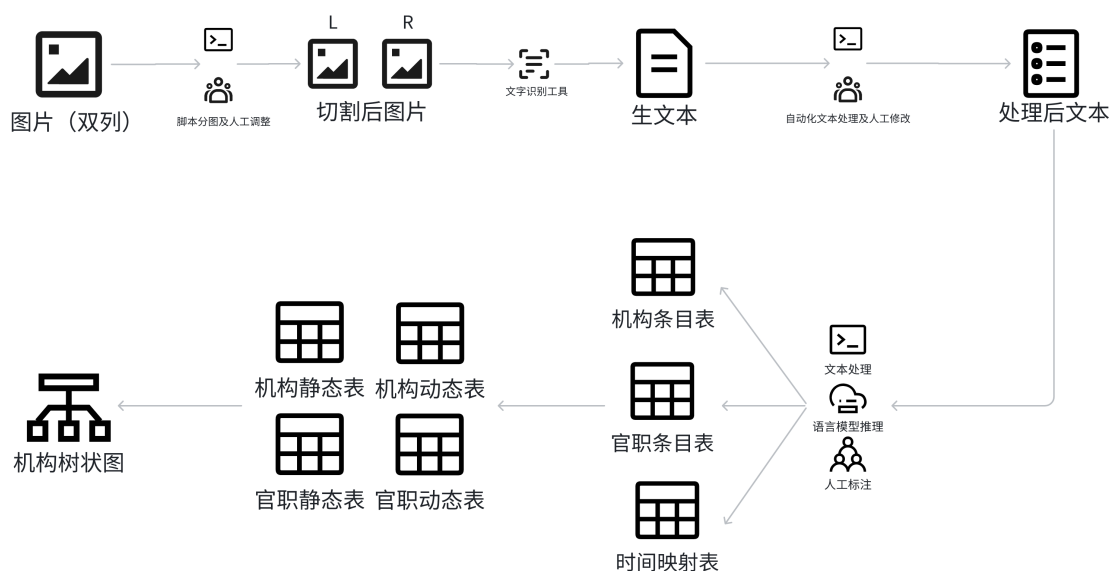


图 4.1 数据处理方案流程图

第五章 官制结构数据可视化

可视化设计的部分需要信息设计领域的知识，也需要通过用户实际体验的反馈来不断进行优化，当前的实现仅进行官制结构树状图的可视化和一些简单的交互性尝试，后续的优化将同历史系和设计方向的专家学者们讨论后再继续进行优化。

需要支持的功能有：

1. 将条目按时间顺序形成事件组成新数据表，并转化为对应的 JSON 格式，为后续的可视化开发库调用提供支持。
2. 需要树状图的支持，能够将当前时间节点下的机构、官职条目都展示出来。
3. 根据节点的颜色等通道对不同特征的节点进行区分。
4. 支持机构节点的收起展开操作。

5.1 数据预处理

因为机构、官职可能存在着不同的存续期，在不同存续期下其所关联的“职掌”、“编制”都会存在变化，经过统计，可以分为以下几类事件：（1）新增（刚开始设置）；（2）取消（罢置）；（3）重启（罢置后重新设置）；（4）变更（职能、编制、官品等信息更新）；（5）移置（隶属情况变更）；（6）合并（多个机构/官职合并为一个）；（7）并入（多个机构/官职由其中一个吸收）；（8）打散（某一机构/官职拆分成多个机构/官职）；（9）拆分（某一机构/官职的部分职能分裂为一个新的下级机构/官职的职能）。

在这九类事件中，部分实现逻辑是类似的，如“变更”与“移置”，都是某一动态字段在新的存续期需要动态更新，“合并”与“并入”都是多个对象被一个对象继承。经过归类，以上九类事件可以由以下五个行为表出：

1. 新增——新增一个机构/官职条目。
2. 变更——已有的数据中的部分条目数据更新。
3. 取消——删除已有的机构/官职条目。
4. 合并——多个机构/官职合并为一个机构/官职，将其职能进行合并拼接赋给新的机构。
5. 打散——原机构/官职信息删除，添加打散后的机构/官职信息（仍然是“新增”与“取消”行为的结合使用，但是因为该行为在语义上是包含两个行为过程的，所以视作一个独立的行为）。

为了方便区分，原先构建于同一个表中的所有字段，根据是否可能会随着时间改变而发生改变划分为动态字段和静态字段分别存储，再根据需要从动态表中生成对应

时间节点的层次数据后再进行内连接（inner join）生成所需数据。

确定了行为和数据集的构造后，接着定义不同事件行为的实现逻辑：

1. 新增：属于此行为的事件包括“新增”、“重启”和“拆分”，这三类事件都属于新的机构从无到有的 0-1 变化，故只需要将对应的数据直接导入即可。
2. 变更：属于此行为的时间包括“变更”和“移置”两项，这类事件属于在已有机构上部分动态字段出现更新，例如“编制”、“职掌”、“上级机构”等，属于 1-1 的映射，故对此的处理逻辑为弹出原数据直接插入新的数据。
3. 取消：该部分只有“取消”一个事件，即将已插入的数据弹出。
4. 合并：该部分包括“合并”与“并入”两个事件，此类事件因为没有显式声明合并后的目标机构的职能的变化，所以在此设计为原机构的所有职能都归于目标机构，所以在原机构数据弹出前，需要将其“职能”汇总，在“新”机构插入时一并插入其“职能”字段中。
5. 打散：该部分只有“打散”一个事件，其行为和“变更”的处理逻辑类似，但是“变更”在语义上属于部分字段的变化，存在着潜在的优化空间，故将“打散”和“变更”两个行为独立开来。

在所需数据导入完成后，根据机构和官职数据的“上级机构”、“隶属机构”字段生成层次数据结构——导出为 D3 开发库所需的 JSON 结构数据，再由开发库中所含的接口进行读取调用。

具体逻辑参看伪代码“Algorithm 1: JSON 数据生成——以机构数据为例”。

5.2 树状图实现

树状图实现直接参考了 D3 可视化开发库的实例，在此简单描述一下这部分主要的代码逻辑以对 D3 开发库所支持的功能有一个简单的了解。

样式设计方面就见仁见智了，本部分工作基于实例的样式进行了有限的调整，后续的改进需要同设计方面的专家一起配合讨论完成。

至于实现部分，因为所涉及的图表过于庞大，实例中并不包含缩放的配置，所以在当前的数据下，对于画布设置了一个比较大的参数，以使得节点的完全展开排布不会重叠。

其次，便是直接调用 D3 开发库中内置的树结构构造函数 (`d3.tree().size([height, width])`) 生成一个树结构对象以存储数据。因为原实例的数据是直接内嵌在代码实现中的，而本次任务需要调用外部的文件，D3 开发库正好支持处理 JSON 格式数据的异步调用 (`d3.json(filename, function)`) 导入 JSON 格式数据，再在 `d3.json` 函数内继续调用 `d3.hierarchy(treeData, function)` 函数转化层次数据——包含父节点 (parent)、子节点

Algorithm 1: JSON 数据生成——以机构数据为例**Input:** 机构动态表, 机构静态表, institutions-dict, specifed-time**Output:**

```

1  读取动态表;
2  for 动态表中条目 do
3      if “变更类型”为“新增”或“重置” then
4          | step 1: 插入新机构条目项 institutions-dict[机构名][开始时间];
5      else
6          if “变更类型”为“变更”或“移置” then
7              | step 2: 根据事件的“开始时间”, 找到 institutions-dict[原机构] 列表
8              | 中最后一个对象补齐“结束时间”;
9              | step 1;
10         else
11             if “变更类型”为“取消” then
12                 | step 2;
13             else
14                 if “变更类型”为“合并”或“并入” then
15                     | 找到“原机构”对应列表中最新的对象, 汇集相应的“职
16                     | 掌”文本数据;
17                     | 对所有“原机构”执行 step 2;
18                     | step 1;
19                     | 将汇集的“职掌”拼接到“新机构”的“职掌”中;
20                 else
21                     if “变更类型”为“打散” then
22                         | // 当前设计下“打散”与“变更”行为一致
23                         | step 2;
24                         | step 1;
25                     else
26                         | // 异常数据
27                     end
28                 end
29             end
30         end
31     end
32 end
33 根据 specifed-time 遍历 institutions-dict 中所有对象, 找到存续期在
    specifed-time 上的对象存入临时机构表 show-ins-dict;
34 读取机构静态表;
35 导入 show-ins-dict 中的静态字段数据;
    // 建树
36 根据每个对象的 parent 字段找到对应的父节点, 将本机构名插入父节点
    children-list 列表中, 并记录根节点 (parent 为空的节点);
37 从根节点开始根据 children-list 递归插入子对象到 children 列表中;
38 将根节点字典转化为 JSON 格式;

```


(children)、高度 (height)、深度 (depth) 的信息。实例的设计将根节点依据高度信息置于画布左侧中间的位置，正好符合官制结构的展示需求，故在这里没有进行修改。

接下来的实现基本没有进行太多修改，但因为是实现相关，也作为对于 D3 开发库的相关支持的介绍，在此简单描述一下生成逻辑。

本部分仍保留在 d3.json 的异步调用函数中，在完成层次数据导入和根节点位置设置后，则进行剩余节点的更新：

1. 调用前述的树结构对象将数据映射，并调用 `descendants()` 函数取得后代节点的拓扑排序集合，构建节点索引。
2. 构建 `enter`、`update`、`exit` 操作集分别处理节点新增、节点更新、节点删除几个事件，这部分是 D3 中比较重要的概念 (如图 5.1)，它处理的是数据集和选择集的对应问题，对于给定数据，如果节点集合有对应的节点，则执行 `update()` 操作 (`update` 没有特殊的添加删除逻辑，一般只补充过渡动画，其余行为和 `enter` 集合的节点特征设置逻辑一致)；若数据集大于选择集，那么数据集中超出的节点部分将执行 `enter()` 操作，将节点插入 `group` 中并设置节点的形状、文本等信息；若数据集小于选择集，则有节点需要删除，则通过 `exit` 集进行 `remove()` 操作——基本的逻辑便是如此，在这一部分，因为需要加上节点名称的信息，而且本实例为树状图实现，所有原实例在 `enter` 和 `update` 集中加入了根据是否存在子节点而选择将节点名称排列在节点的左侧或右侧的判断逻辑。

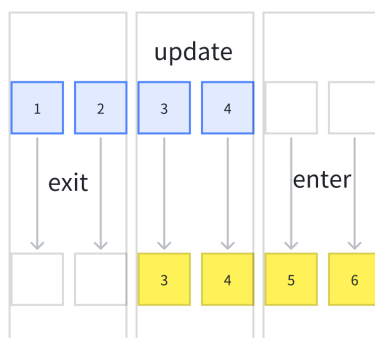


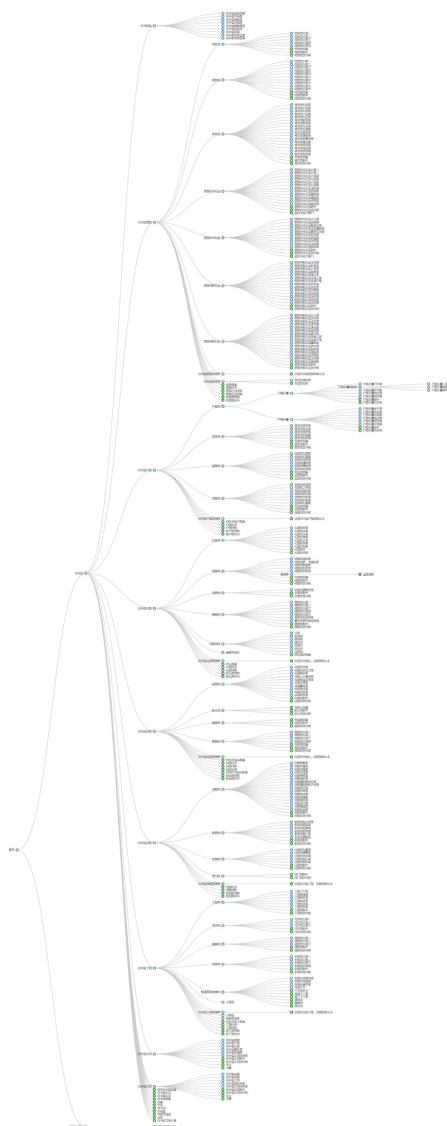
图 5.1 `enter`、`update`、`exit` 集

3. 对于边操作也进行同样设置，但是需要添加边的笔触路径，此路径由官方给出的贝塞尔曲线路径表达。

修改的部分在于：

1. 重新实现了数据读取的方式：重新实现外部 JSON 文件的读取。
2. 根据体验效果修改了画布大小及节点的部分特征设置。

最终效果如图 5.2。

图 5.2 官制结构树状图实现^①

5.3 机构、官职节点区分

机构和官职节点属于两类不同的节点，原始设计在插入过程中并没有对这两类节点对象进行区分，而是视为了同一类节点。为了区分两类节点对象，分别在 `enter` 以及 `update` 事件中添加判断：根据其“类别名”字段是否为“机构”而分配不同的颜色的逻辑，以将机构和官职节点进行区分。后续若需要更细致化的区分如官、吏的节点的区分，则基于此逻辑进行扩展即可。

效果见图 5.3。

^① 以部分中央机构——元丰改制后尚书省的数据为例，下同。

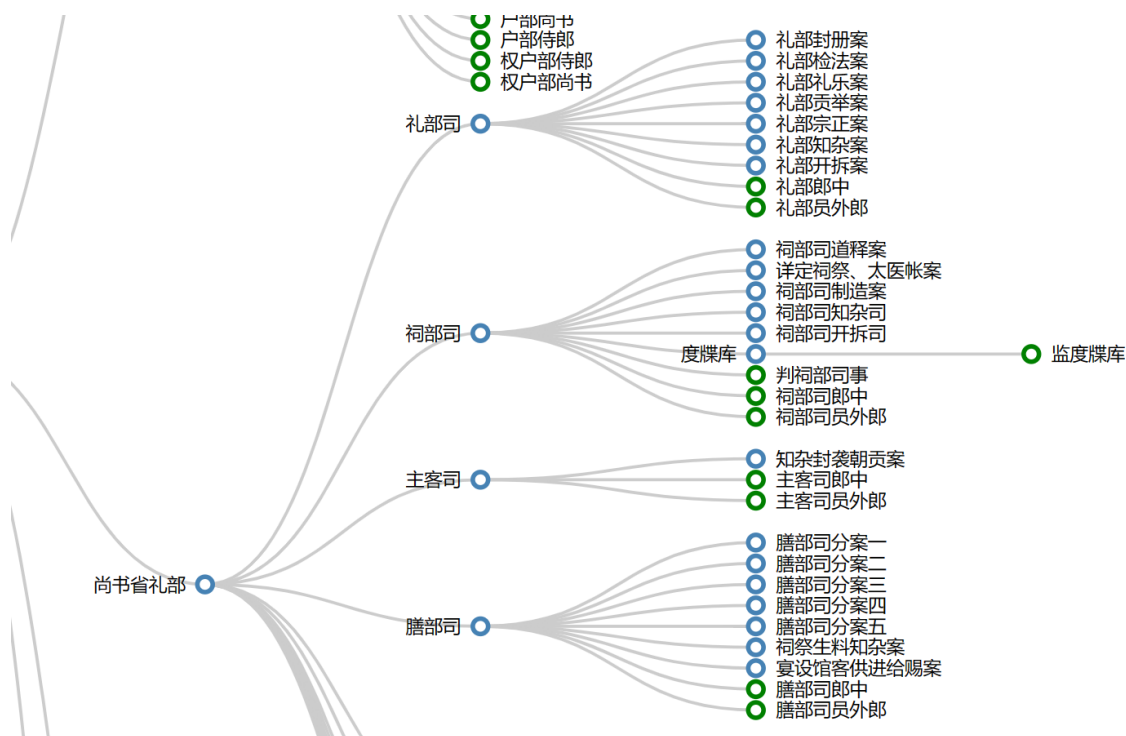


图 5.3 机构、官职节点的节点特征区分

5.4 机构、官职节点的折叠和展开

元丰改制后的中央机构呈现出一个扁平的矮树的层次，经过讨论，在使用过程中未必需要将所有的数据都完整地进行呈现，用户往往只想要关注某部分的数据，其他数据进行折叠可以很大程度地改善页面的利用率。

因此，在前面的设计之上，在 `enter()` 操作上关联鼠标左键单击事件，负责触发节点的“折叠”、“展开”的切换行为。实现上，新增 `_children` 保存隐藏的子节点信息，根据节点的状态在点击事件发生后切换 `children` 与 `_children` 字段来实现节点的“折叠”与“展开”，点击事件触发完成字段切换后重新调用 `update()` 操作进行更新。其中，保存节点位置信息的参数有 `x0,y0,x,y` 四个，其过渡动画由 `x0,y0` 与 `x,y` 的移动行为捕捉，而在每次 `update` 行为的最后，将 `x,y` 信息赋给 `x0,y0` 而实现节点的旧位置更新。

效果如图 5.4。

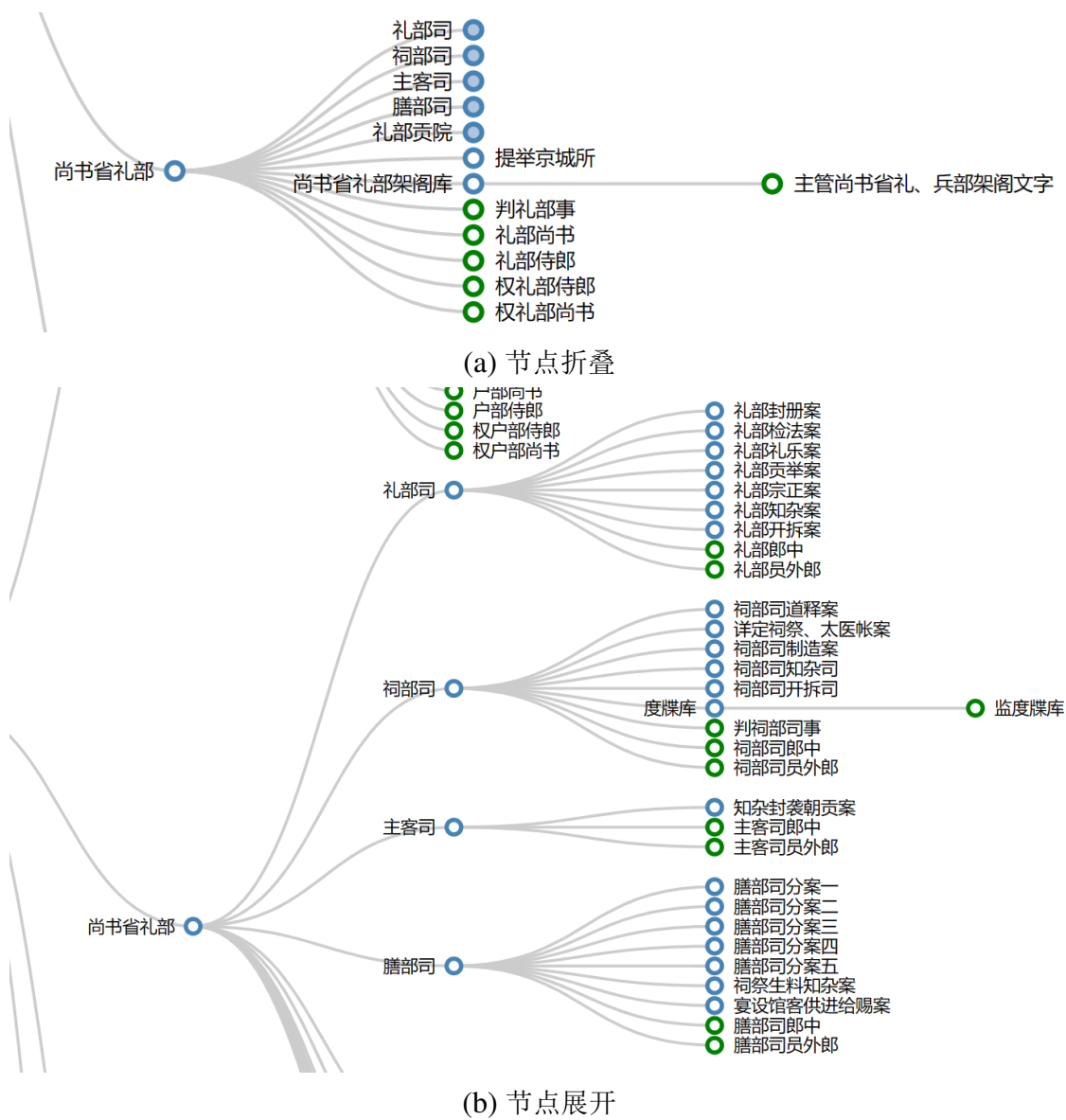


图 5.4 节点的展开与折叠示例

第六章 总结和展望

6.1 工作内容回顾

回顾本次工作，其意义在于以下几个方面：

1. 补充了宋代官制机构可视化方面存在的需求：至今为止，在可视化方面上，已经有了不少关于职业流动、人物分布、战场形式等的可视化及分析工作，但是官制结构作为职业流动的其中一种呈现形式，目前暂时还未找到一个相关的可视化实现。本工作补充了官制结构可视化的实现，为完善进一步的探索工作——与前述的包括职业流动、人物流动等可视化功能的结合，增加可探索的维度——奠定了基础，提供了一种全新的呈现形式^[50]。对于历史学习者来说，这种形式也将是一种非常直观、形象的官制结构的表达，使得使用者从中既能对整体的官制情况进行概览，也能通过节点的交互对其中更详细的信息进行学习。
2. 在本次工作中，主要的工作重心被放在了数字化的部分，“辞典”的数据量仍然十分巨大，本次工作覆盖的范围仅仅是其中的一小部分。从前述的流程中可见，只靠人工手段进行数字化处理和官制结构的生成，是一项非常费时费力的过程。本次工作从初始的纯人工提取尝试开始，到后续的逐步向半自动化流程转化，目前的效率相较于开始工作初期的平均工作效率已经提升了 1-3 倍，为加快后续的宋代官制结构数据的生成提供了基础。但是当前的流程平均的处理时间仍然存在着很大的优化空间，未来将会结合大语言模型的测试进一步地减少人工处理的工作量，以期最后达到数据转换工作效率的最大化。
3. “辞典”中存在着一定程度的事实冲突和缺失，在历史文献研究过程中是不可避免的。因为文献资料的不同或者记录的详细程度的差异，不同文献所表达的历史事实未必是完全一致的。由于层次化数据生成的需求，其数字化到可视化的流程亦可以作为数据检查和校对之用，包括机构的缺失、机构关系的冲突等，在生成对应的 JSON 对象的过程中，此类冲突将会导致层次数据生成的报错，反向则可以作为“辞典”甚至更多材料的事实冲突的一种检查工具。后续随着更多字段数据如“官品”、“编制人数”等和其他功能的结合，将会继续强化其数据检查的工具的地位。
4. 从某种意义上讲，因应不同的可视化需求，不同可视化系统的实现逻辑都是需要进行个性化定制的，在实现上没有一种可以直接照搬的公式可以套用。但是从可视化广义的视角下，不考虑具体的实现细节，整个可视化设计的流程，仍然是有着一定的可以遵循的路径的，亦或是，作为一个毕业设计项目，其本身就是需

要这么一个完整的工作路径，本次工作也是为了对这一流程进行探索，为日后的研究或者设计形成一套严谨有效的方法论，即：背景调研-需求讨论-数据结构设计及优化-数据提取-数据处理-数据集实现-系统设计与实现-体验与反馈收集-系统优化，这一工作流的循环^[51]。

6.2 不足和提升方向讨论

截止至目前的工作，仍然只是一个基本的雏形，在数字化和可视化的工作上，依然存在着很大的提升空间，接下来将分述之：

数据预处理部分仍然有很大比例的工作需要人工参与，包括图像数据数字化、文本数据结构化：

1. 简单的从中间切分图片的逻辑显然并不是一个很好的图像切割方案，每每需要人工检查图像切割的质量进行重新调整。在调研了 OCR 工具的工作流程后，或许可以通过图像处理的技术对根据列间距进行定位和切割。
2. 大语言模型的强大功能在本次工作中并没有得到很好的挖掘，存在着巨大的潜力，在文本检查、文本数据转表格数据的工作中，通过提供筛选好的语料数据和示例以及提示词的优化，可能带来很大的效果的提升。后续将会继续进行相关方面的测试，逐步向全流程自动化的目标靠近。

可视化工作上亦有很大的修改空间。本次工作的可视化实现暂时只完成了在不同时间节点上的树状结构呈现、机构/官职类型数据的节点颜色区分以及节点的“折叠”与“展开”三个需求，在以下几个方面仍可进行改进：

1. 布局设计：本次工作只是基于开发库所提供的实例进行了简单的样式修改，虽然能够满足呈现官制层次结构的要求，但是在页面的美观度上就略显不足了，相关的工作将会与设计方向的老师同学们进行讨论，结合历史系方向老师同学们的使用体验进行相应的页面布局的优化。
2. 交互设计：本次设计只实现了节点的展开与收缩以及官职、机构节点的区分，并没有考虑到多视图的可能性，如纯机构/官职节点的展示、一级节点展开或部分节点展开等功能，使之适应更多的探索需求。
3. 节点连接形式：在最新的数据检查中，发现条目数据并不是简单的层次数据结构，因为存在多个机构共同管理某一下级机构的情况，当前简单的树状图只能通过增加冗余节点来实现机构间隶属关系的展示，但并不能体现其机构共管的特点。仍然需要调研其他可选的实现，或基于图的实现来作为新的节点连接形式的尝试，找到该类节点特性的表现形式。

数字化、可视化工作是一项繁杂且涉及大量领域知识的工作，它似乎可以作为一

项独立的学科，但又不得不与大量的其他领域的知识交融从而落地^[52]。在这个方向上，有着大量的研究者在前赴后继地为着更好的可视化效果努力，为着帮助更多学科学习和研究做出贡献^[53]。无论在研究价值还是实用价值上，都存在着巨大的发展潜力，亟待更多相关方面的专家学者们加入，增加其覆盖的广度和深度^[54]。

参考文献

- [1] 龚延明. 宋代官制辞典（增补本）[M]. 北京: 中华书局, 2001.
- [2] 鄧小南. 数字人文與中國歷史研究[J]. 中国文化, 2021(01): 11-14.
- [3] KNOWLES A. Emerging trends in historical GIS[J]., 2005, 33: 7-13.
- [4] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报, 2016, 42(2): 15.
- [5] UNSWORTH J. What is Humanities Computing, and What is Not?[J]. Paderborn: mentis, 2002.
- [6] 王晓光. “数字人文”的产生、发展与前沿[C]//方法创新与哲学社会科学. 武汉: 武汉大学出版社, 2010.
- [7] 郭英剑. 数字人文: 概念、历史、现状及其在文学研究中的应用[J]. 江海学刊, 2018(3): 8.
- [8] 朱本军, 聂华. 跨界与融合: 全球视野下的数字人文——首届北京大学“数字人文论坛”会议综述[J]. 大学图书馆学报, 2016, 34(005): 16-21.
- [9] 刘炜, 谢蓉, 张磊, 等. 面向人文研究的国家数据基础设施建设[J]. 中国图书馆学报, 2016, 42(5): 11.
- [10] 龚延明. 《宋代登科总录》与创新的宋代精英数据库[J]. 浙江大学学报(人文社会科学版), 2017, 47(01): 29-34.
- [11] Harvard University, Academia Sinica, Peking University. China Biographical Database (CBDB)[EB/OL]. 2019. <https://projects.iq.harvard.edu/cbdb>.
- [12] FULLER M, WANG H. Structuring, Recording, and Analyzing Historical Networks in the China Biographical Database[J/OL]. Journal of Historical Network Research, 2022, 5(1). <http://jhnr.uni.lu/index.php/jhnr/article/view/test>. DOI: 10.25517/jhnr.v5i1.123.
- [13] KNOWLES A. Past time, past place: GIS for history[M]. [S.l. : s.n.], 2002.
- [14] GREGORY I N, ELL P S. Historical GIS: Technologies, Methodologies, and Scholarship[M]. [S.l.]: Cambridge University Press, 2007.
- [15] 马继. 动态视图在文字编管理系统中的应用研究[J]. 数字人文, 2023(03): 77-87.
- [16] 张翰兴. 基于序参量理论的三国鼎立成因可视化设计研究[D]. 哈尔滨工业大学, 2018.
- [17] STALEY D. Historical Visualizations[J]., 2000, 3.
- [18] 王兆鹏. 从“年谱”到“编年系地谱”——重建作家年谱的理念与范式[J]. 文学评论, 2021(02): 17-24.
- [19] 王兆鹏, 邵大为. 数字人文在古代文学研究中的初步实践及学术意义[J]. 中国社会科学, 2020(08): 108-129+206-207.
- [20] KORENGOLD A. REVEALING DATA: VISUALIZATIONS IN HISTORICAL COLLECTIONS[EB/OL]. 2023. <https://circulatingnow.nlm.nih.gov/2023/07/13/revealing-data-visualizations-in-historical-collections/comment-page-1/>.

- [21] STALPH F, HERAVI B. Exploring Data Visualisations: An Analytical Framework Based on Dimensional Components of Data Artefacts in Journalism[J/OL]. Digital Journalism, 2023, 11(9): 1641-1663. eprint: <https://doi.org/10.1080/21670811.2021.1957965>. <https://doi.org/10.1080/21670811.2021.1957965>. DOI: 10.1080/21670811.2021.1957965.
- [22] GOODCHILD M F. Communicating Geographic Information in a Digital Age[J/OL]. Annals of the Association of American Geographers, 2000, 90(2): 344-355. eprint: <https://doi.org/10.1111/0004-5608.00198>. <https://doi.org/10.1111/0004-5608.00198>. DOI: 10.1111/0004-5608.00198.
- [23] 位通, 桑宇辰, 史睿. 基于知识重构的年谱时空可视化呈现——以《朱熹年谱长编》为例[J]. 中国图书馆学报,
- [24] 胡静. 文本挖掘与码库思: 以朝鲜译官在清朝的贸易网络为例[J]. 数字人文, 2023(03): 30-49.
- [25] 赵薇. 新时代 | 数字时代人文学研究的变革与超越——数字人文在中国[J/OL]. 探索与争鸣, 2021, 1(6), 191: 191-206. <http://www.tsyzm.com/CN/Y2021/V1/I6/191>.
- [26] 史睿. 索引与知识发现[J]. 中国索引, 2006, 4(1): 8.
- [27] 邓柯, 包弼德, J.LI K, 等. 专门领域中文文本的无监督分析[J]. 数字人文, 2023(03): 13-29.
- [28] 李绅, 胡韧奋, 诸雨辰. 古籍标点与专名的智能识别技术研究[J]. 数字人文, 2023(03): 63-76.
- [29] 严丹, 唐蓓, 黄磊. 国外数字人文教科书中的中国案例及其对中国形象的构建[J]. 数字人文, 2023(03): 160-176.
- [30] 于亚秀, 李欣. 数字人文视域中的古籍文本标注方法研究——以 MARKUS 为例[J]. 大数据, 2022, 8(6): 11.
- [31] 李斌, 王璐, 陈小荷, 等. 数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J]. 大学图书馆学报, 2020, 38(5): 10.
- [32] 李林芳, 杨海峥, 袁晓如. 古籍流传的可视化[J]. 图书馆论坛, 1-9.
- [33] 北京大学可视化与可视分析实验室. 可视化看中国[EB/OL]. 2024. <https://vis.pku.edu.cn/vis4china/>.
- [34] CHO I, DOU W, WANG D X, et al. VAIroma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 210-219. DOI: 10.1109/TVCG.2015.2467971.
- [35] 龚延明. 宋代中书省机构及其演变考述[J]. 杭州大学学报 (哲学社会科学版), 1990(03): 43-50+67.
- [36] 龚延明. 宋代中央机构剖析[J]. 浙江学刊, 1993(03): 113-117. DOI: 10.16235/j.cnki.33-1005/c.1993.03.026.
- [37] 史睿. 论中国古籍的数字化与人文学术研究[J]. 北京图书馆馆刊, 1999(02): 28-35.
- [38] 夏翠娟. 中国历史地理数据在图书馆数字人文项目中的开放应用研究[J]. 中国图书馆学报, 2017, 43(2): 14.
- [39] CHEN G, LI M, CHEN D. Compiling of thematic map in historical geography studies for Jiankang in the period of the Six Dynasties (220 589 A.D.)[C/OL]//LI M, WANG J. Geoinformatics 2007: Cartographic Theory and Models: vol. 6751. [S.l.]: SPIE, 2007: 67511G. <https://doi.org/10.1117/12.760100>. DOI: 10.1117/12.760100.

- [40] WANG Y, LIANG H, SHU X, et al. Interactive Visual Exploration of Longitudinal Historical Career Mobility Data[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(10): 3441-3455. DOI: 10.1109/TVCG.2021.3067200.
- [41] 严承希, 王军. 数字人文视角: 基于符号分析法的宋代政治网络可视化研究[J]. 中国图书馆学报, 2018, 44(5): 17.
- [42] 张静. 数字人文中历史人物数据的可视化应用研究[J]. 湖南大学,
- [43] 王兆鹏, 蒋晓晓. 时空一体化——唐宋文学编年地图平台的学术理念与学术价值[J]. 三峡论坛(三峡文学·理论版), 2020(05): 20-27+35.
- [44] LIU J, WAN Z. The Making of a Sacred Landscape: Visualizing Hangzhou Buddhist Culture via Geoparsing a Local Gazetteer the Xianchun Lin' an zhi 咸淳臨安志[J/OL]. Religions, 2022, 13(8). <https://www.mdpi.com/2077-1444/13/8/711>. DOI: 10.3390/rel13080711.
- [45] 张昕, 袁晓如. 树图可视化[J]. 计算机辅助设计与图形学学报, 2012, 24(09): 1113-1124.
- [46] YOUSSEF D, YUTA Y, TAKUJI Y. FuncTree2: an interactive radial tree for functional hierarchies and omics data visualization[J]. Bioinformatics, (21): 21.
- [47] MUNZNER T, SAFARI A O M C. Visualization Analysis and Design[M/OL]. [S.l.]: A K Peters/CRC Press, 2015. <https://books.google.com/books?id=FUJRzQEACAAJ>.
- [48] 龚延明. 唐宋官、职的分与合——关于制度史的动态考察[J]. 历史研究, 2015(05): 92-106+192.
- [49] 骆翠欣, 谭红丽. 字体强调对语言理解的影响及其机制[J]. 心理学进展, 2022, 12(1): 193-199.
- [50] 邓小南. 走向“活”的制度史——以宋代官僚政治制度史研究为例的点滴思考[J]. 浙江学刊, 2003(03): 98-102. DOI: 10.16235/j.cnki.33-1005/c.2003.03.014.
- [51] 程佳羽, 史睿. 古籍数字资源的知识库建设解析[J]. 数字图书馆论坛, 2006(12): 5.
- [52] 威拉德·麦卡蒂, 韩玉凤. 作为跨学科的人文计算[J]. 数字人文, 2023(03): 50-62.
- [53] 袁晓如. 可视化多学科设计实践[J]. 世界建筑, 2022(11): 112-113. DOI: 10.16414/j.wa.2022.11.017.
- [54] 袁晓如, 张昕, 肖何, 等. 可视化研究前沿及展望[J]. 科研信息化技术与应用, 2011, 2(04): 3-13.

致谢

首先衷心感谢我的指导老师，也是本学期进行了两周的可视化课程的任课老师袁晓如老师，在选择这个题目的时候，我对这方面的知识了解得非常少，属于是究极半路出家的类型了。还记得当时抱着了解一下的目的和老师讨论结束时，老师问到“你对这个感兴趣吗？你觉得这个有意思吗？要不要一起参与进来的时候？”时，在大学的生活中慢慢已经磨灭的仅仅为了“有意思”而去投入时间去学习去解决问题的那种朴素的欲望，又重新被点燃了，感谢老师指引我找到感兴趣的研究方向。

尽管在过程中走了很多弯路，袁老师仍然给予了耐心的等待，由于本科科研经历的空白，我对于整个工作流程的认识一直处在一个十分混乱的状态，为此没少拖慢项目整体推进的速度。在这个过程中，是老师在包容我不断碰壁的情况下为我点出问题和解决方向，帮助我从基本零了解的状态下开始有序推进我的毕业论文工作，直至到此顺利完成全文。“做这个毕业论文也是你本科生的一次科研的经历”，当时听到老师这么说的時候体会还不深刻，现在也算多多少少明白其中的酸甜苦辣了，感谢老师的帮助，补全了本科生涯险些没有任何科研经历的遗憾。

此外，还要感谢“宋代人物研究”小组的老师、师兄和同学们——史睿老师、高柯立老师、马驰腾师兄、蒋瑞珂师兄、龙天龔同学、张毓祺师兄、张佳帝师兄——的帮助。感谢两位老师在历史问题方面的耐心解答，感谢驰腾师兄在这个工作过程中对于实现上的各种大大小小的问题的解惑，感谢瑞珂师兄、天龔同学、毓祺师兄、佳帝师兄的支持和努力，希望最后整个项目可以做的越来越好！

感谢鳄梨科技的领导们的理解和批假，让我能从实习工作中抽身出来全力投身到毕业论文的工作中去。

最后，就是要感谢一直在支持我的家人、同学和朋友们（特别是元培男篮和104宿舍的各位），在因为工作变得越来越自闭的情况下，感谢你们的关心和支持，要感谢的地方太多了，也不知该从哪里开始罗列，反正从生活上、学习中都得到了你们巨大的支持，若是一一罗列恐怕就有水字数的嫌疑了，就允许我将之放在心里不在此一一详述了吧。除了学生这个身份，我们还是共同在这个社会中生存的人，因为有了你们，这个社会纽带才能一直都在！

最后的最后，再次感谢几位老师和同学们，这次学科交叉的经历真的十分难忘！